

UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Prediction of cancer phenotypes through the integration of multi-omic data and prior information

Permalink

<https://escholarship.org/uc/item/5cs2x2bz>

Author

Uzunangelov, Vladislav Jordanov

Publication Date

2019

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike License, available at <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

**PREDICTION OF CANCER PHENOTYPES THROUGH THE
INTEGRATION OF MULTI-OMIC DATA AND PRIOR
INFORMATION**

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Vladislav J. Uzunangelov

June 2019

The Dissertation of
Vladislav J. Uzunangelov
is approved:

Professor Joshua Stuart, Chair

Professor David Draper

Professor Christopher Benz

Dean Lori Kletzer
Vice Provost and Dean of Graduate Studies

Copyright © by
Vladislav J. Uzunangelov
2019

Table of Contents

List of Figures	v
List of Tables	xvii
List of Algorithms	xviii
Abstract	xix
Dedication	xxi
Acknowledgments	xxii
1 Introduction	1
2 Hierarchical VIPER (hVIPER)	5
2.1 Uveal Melanoma	6
2.1.1 Introduction	6
2.1.2 Results	8
2.1.3 Data	14
2.1.4 Methods	20
2.2 Castration-Resistant Prostate Cancer	27
2.2.1 Introduction	27
2.2.2 Results	28
2.2.3 Methods	36
2.3 Conclusion	43
3 Gene Essentiality Prediction via Multiple Pathway Learning	44
3.1 Introduction	44
3.2 Results	48
3.2.1 Overall Results - Subchallenge 1	48

3.2.2	Results - Multiple Pathway Learning	48
3.3	Data	62
3.4	Methods	62
3.4.1	Random Forest	62
3.4.2	Multiple Pathway Learning (MPL)	63
3.4.3	MPL-RF Ensemble	64
3.4.4	Kernel construction	64
3.4.5	Hyperparameter selection	68
3.5	Conclusion	71
4	AKLIMATE: Algorithm for Kernel Learning with Approximating Tree Ensembles	72
4.1	Introduction	72
4.2	Results	77
4.2.1	Microsatellite Instability	77
4.2.2	Breast Cancer Survival	84
4.2.3	shRNA knockdown viability	89
4.3	Discussion	105
4.4	Data Acquisition and AKLIMATE pre-processing	112
4.4.1	Microsatellite Instability	113
4.4.2	METABRIC Survival	113
4.4.3	shRNA Knockdown Profiles	115
4.4.4	Gene Set Collections	116
4.5	Methods	117
4.5.1	Random Forest	118
4.5.2	Kernel Learning	119
4.5.3	Multiple Kernel Learning	120
4.5.4	RF Kernel	122
4.5.5	Stacked Learning	126
4.5.6	AKLIMATE as a Super Learner	128
4.5.7	AKLIMATE algorithm	130
4.5.8	Feature and feature set weights	133
4.5.9	AKLIMATE hyperparameter settings	134
4.5.10	Implementation	136
5	Future Directions	137
5.1	Multiclass AKLIMATE	137
5.1.1	Minimal Feature Predictors of TCGA Subtypes	138
5.2	Non-linear Stacked Learners	150
5.3	Logical Dependency Features	150
5.4	Pathway-Pathway Dependency Map	151

List of Figures

2.1	Clustering solutions based on A) mRNA and B) lncRNA features. Covariate annotation tracks show selected genomic and clinical features. The lower heatmap displays the expression profiles of 12 genes used in a prognostic test for the risk of developing metastasis [60], with blue text highlighting genes that are on chromosome 3. High-risk primary tumors show low expression of eight of these genes and high expression of four genes (yellow versus green panels at the left). BAP1 structural alterations that include alternative splicing and rearrangements were detected by assembly of RNA-seq and DNA-seq data. Leukocyte fraction was estimated from DNA methylation data. LOH, loss of heterozygosity. *, **, *** p value < 0.1, 0.01, 10 ⁻⁵ , Fisher's Exact or Chi-square test. C) Kaplan-Meier curves for overall and metastasis-free survival for mRNA and lncRNA subtypes, with log-rank p-values. Figure panels taken from [117].	11
-----	---	----

2.2	Significant findings from the A) PARADIGM and B) hVIPER pathway biomarker analysis (see 2.1.4.3). A) The heatmap displays scaled mean IPLs of regulatory hubs that have at least 10 downstream targets and are differentially active in at least one of the one-vs-all-other cluster comparisons. The panel at the left identifies the clusters in which a hub is differentially active (yellow). B) The heatmap displays master regulators (MRs) that have differential activity levels in at least one of the one-vs-all-other cluster comparisons. The panel at the left identify the lncRNA cluster in which MRs are differentially active (yellow). For each cluster, gene biomarkers with similar findings across the two methods are highlighted in boldface. The similarity can be based on: 1. complexes or families with components identified by both methods, 2. genes within the same pathway showing complementary inferred activity, or 3. abstract processes linked to any of the above. Figure panels taken from [117].	16
2.3	A) Correlation network for lncRNA clusters 3, showing PARADIGM pathway features, hVIPER regulators, and lncRNAs. Red and blue edges indicate Spearman correlations ($ \rho > 0.5$) between the expression of a differentially expressed lncRNA and the inferred activity of a differentially active PARADIGM or hVIPER feature. The color of each node reflects differential expression for a lncRNA, and relative activity for a PARADIGM/hVIPER feature (red for overexpressed/active, blue for underexpressed/inactive). B) A subnetwork highlighting the lncRNAs with high degree of correlation to regulators in the proliferation and DDR pathways. Both panels generated by myself, with input from Christina Yau and Gordon Robertson, for the TCGA UM working group. Panel A) also appears in [117].	18

2.4	A) Correlation network for lncRNA clusters 4, showing PARADIGM pathway features, hVIPER regulators, and lncRNAs. Red and blue edges indicate Spearman correlations ($ \rho > 0.5$) between the expression of a differentially expressed lncRNA and inferred activity of a differentially active PARADIGM or hVIPER feature. The color of each node reflects differential expression for a lncRNA, and relative activity for a PARADIGM/hVIPER feature (red for over-expressed/active, blue for underexpressed/inactive). B) A subnetwork highlighting the breadth of associations of LINC00152 and BANC1 with regulators in the immune response, DDR, MYC/MAX and hypoxia signaling pathways. Both panels generated by myself, with input from Christina Yau and Gordon Robertson, for the TCGA UM working group. Panel A) also appears in [117].	20
2.5	Hierarchical VIPER for the imputation of kinase activities. ssVIPER is performed for all TFs and kinases with sufficient number of targets in their regulons (bottom level). The inferred activities are used to rank regulators based on difference in mean activity across the dichotomy of interest. Standard VIPER analysis is then run with the kinase regulome and the regulator ranking to produce the final estimates for kinase activities (top level).	23
2.6	Site of biopsy for CRPC patients.	29
2.7	Concordance between the WCDT AR signature and the Mendiratta signature. Each point is a combination of differential (post- minus pre-treatment) scores for each of the two signatures.	31
2.8	Significant Master Regulators (FDR<0.2) discovered by hVIPER in paired-biopsy CRPC patients. NES is the differential activity score between pre- and post-treatment biopsies - positive scores (red) indicate higher differential activity in post-treatment biopsies, while negative scores (blue) correspond to relative inactivation.	32

2.9	Full network of significantly differential Master Regulators. Nodes are colored by differential activity as in Fig.2.8. Shapes represent types of MR. Groups identified by MR Combinatorial Analysis are shown as multi-regulator nodes. Links are colored by type of interaction - green for transcriptional regulation; purple for post-translational regulation. Links to and from a module indicate regulatory interactions present between any of the individual module members and another MR/module. A module regulatory link is thicker if multiple module members interact with another MR/-module - e.g. connecting a two-member module with a three-member could produce a maximum of six links among the Cartesian product of module members. If in fact only three of these are present in the MR regulome, the link thickness will be proportionately decreased. All links are plotted on a common thickness scale, with the thinnest lines corresponding to links between individual MRs.	33
2.10	Main biological processes influenced by members of the four Regulator Combinatorial Modules characterizing pre- vs post- treatment CRPC (see section 2.2.3.2). Each regulator is involved in one or more processes as indicated by the regulator's coloring. Modules are shown as groups of multiple regulators.	35
2.11	CREB1 RNAi and 666-15 experiments. A) CREB1 RNAi suppresses enzalutamide-resistant cell growth. The indicated cell lines were transfected with CREB1 RNAi or NTC RNAi. Cell viability was determined 120 hours later using fluorescent-labeled cell imaging. Protein lysates were harvested at 120 hours and probed with the indicated antibodies. B) CREB1 inhibitor 666-15 suppresses enzalutamide-resistant cell growth. The indicated cell lines were treated with dose escalation of the CREB1 inhibitor 666-15. Cell viability was quantified 72 hours later using fluorescently-labeled cell imaging. C) 666-15 treatment blocks the transcriptional function of CREB1. The indicated cell lines were transfected with a CRE-luciferase reporter plasmid. Cells were treated with 300nM of 666-15 or vehicle 24 hours post-transfection, and 10 μ M forskolin added 30 minutes post-treatment. Cells were harvested 6 hours post-treatment, and luciferase signal was quantified and normalized to total protein concentration. Experiments and data analysis performed by Joshi Alumkal.	37

2.12	Balanced permutation procedure used in CRPC paired biopsy analysis. The true background model must exclude same-patient biopsies from matching up in the permuted folds.	38
3.1	Schematic of the Broad-DREAM gene essentiality prediction challenge. The challenge consisted of three related sub-challenges. Sub-challenge 1 addressed building a model that best predicts all gene essentiality scores in the held-out test set, using any feature data. Figure taken from [59].	47
3.2	Team’s performances on DREAM gene essentiality prediction sub-challenge 1. The average Spearman correlation across all predicted genes is shown at the top. Black lines in the middle of blue boxes (“team rank”) display the average rank of each team’s scores across 14,738 genes. Black lines below the boxplots summarize the Wilcoxon signed rank test results by dividing the teams into groups. Each group is denoted by a distinct black line, which contains the teams with no statistically significant difference. The red line displays the rank of an ensemble model constructed from all models with rank less than or equal to the corresponding point on the x axis. Red stars on the x-axis indicate the three joint co-winning teams, determined by non-parametric Friedman’s test - MPL-RF is the middle star. Figure taken from [59].	49
3.3	Mean Spearman correlation for a baseline lasso regression model (red dot) across all prediction tasks. The histogram represents mean Spearman correlations for a null distribution computed by randomly permuting the labels of the test set 10,000 times and applying the baseline model to the permuted test data. Figure taken from [59].	50
3.4	Relationship between number of 6σ measurements in an shRNA profile and the predictive accuracy of trained models for the best performers in subchallenge 1. The curves represent a smoothed approximation by local polynomial regression [27] computed via the R function <i>loess</i> with default parameters.	53
3.5	Spearman correlations for top subchallenge 1 methods on 426 high-confidence shRNA profiles. DREAM9-w1 is the top-ranked method overall in subchallenge 1, MPL-RF is the second best, and DREAM-w2 the third best. P-values for Wilcoxon signed-rank test pairwise comparisons.	54

3.6	Leading edge genes for the top 10 canonical pathways from the GSEA analysis in Table 3.2. Spearman correlations between method predictions and each shRNA knockdown test profile shown for MPL, RF, and the MPL-RF ensemble. Vertical dotted lines mark boundaries between KD pathways. Pathways are ranked from left to right, with the leftmost the most enriched (REACTOME ATP FORMATION). Genes within pathways are ranked by the Spearman correlation of the MPL predictions to the true test set labels.	57
3.7	An overview of PP - KD pathway association strength score derivation, with mRNA Splicing KD pathway as an example. The MPL models for the leading edge gene are queried for their component non-zero weight PPs. If a PP has positive weights for both its expression and copy number kernels, they are summed to produce a composite weight. The PP composite scores are then averaged across leading edge models to give a measure of PP importance for modeling genes in a KD pathway.	58
3.8	PP-KD pathway dependency map for the top 20 KD pathways and the top 20 PPs associated with each of them. Edge weights are computed as described in Fig.3.7. The map is a bipartite graph between KD (green) and PP(red) pathways. The inner circle of PPs contains non-specific ones, i.e. PPs that are among the top predictors for all or almost all KD pathways. Similarly, the outer circle of PPs contains specific ones - PPs that appear among the top predictors of 6 or fewer KD pathways.	61
3.9	Distributions of A) Expression filtered and B) Copy number filtered gene set lengths. Filtering steps as described in text.	67

4.1	AKLIMATE performance on predicting MSI in UCEC TCGA. A) Performance of AKLIMATE and BMKL on classifying MSI-H vs MSI-L+MSS. AUC computed for 50 75%/25% stratified train/test splits. P-values for Wilcoxon signed-rank test pairwise comparisons. Methods: aklimite - aklimite on full collection of feature sets; aklimite-reduced - aklimite with 196 PID pathways; sbmkl - sparse single-task BMKL; dbmkl - dense single-task BMKL; sbmtmkl - sparse multi-task BMKL; dbmtmkl - dense multi-task BMKL. Multi-task BMKL models were trained to simultaneously predict MSI status on UCEC and COADREAD cohorts. B) Top 10 most predictive AKLIMATE feature sets and top 50 most predictive features. Expression of top 50 features (left heatmap); Membership of most predictive features in most predictive feature sets (right heatmap). Features are organized by KNN clustering into 3 groups, followed by hierarchical clustering within each cluster. Feature set model weights scaled to a sum of 1 (barplot, top of right heatmap). Feature model weights scaled to a sum of 1 (Barplot, right of right heatmap). Feature and feature set weights averaged across 50 train/test splits.	80
4.2	AKLIMATE results for the reduced AKLIMATE model (using only 196 PID pathways) on UCEC TCGA cohort. Figure shows top 10 most predictive feature sets and top 50 most predictive features from 50 stratified 75% train/25% test splits. Figure organized as Fig. 4.1.	82
4.3	AKLIMATE performance on predicting MSI-High vs MSI-Low+MSS in TCGA COADREAD cohort. AUC computed for 50 75%/25% stratified train/test splits. P-values for Wilcoxon signed-rank test pairwise comparisons. Methods as in Fig.4.1.	83

4.4 AKLIMATE performance on predicting survival at 2000 days in the METABRIC cohort. A) Performance of AKLIMATE under different data type combinations. EXP+CNV - AKLIMATE with genomic features only; clinical - a RF model run with the clinical variables only; EXP+CNV+CLINICAL - AKLIMATE with genomic features as "local" variables and clinical features as "global" ones. FSMKL and BCC dashed lines show mean performances for the two models under 5-fold cross-validation as shown in [129]. B) AKLIMATE results highlighting the top 10 most predictive feature sets and top 50 most predictive features. Figure organized as Fig. 4.1. Clinical variables shown as column annotations; they are included only if among the top 50 most informative features in the model. Clinical variables are ranked from top to bottom by relative predictive contribution. Survival status is a binary variable representing survival at 2000 days (labels) while days survived shows actual duration of survival. Samples sorted by days survived within the two classes. Feature and feature set weights averaged across 50 train/test splits. 87

4.5	Prediction of cell line viability after shRNA gene knockdowns. A) RMSEs of AKLIMATE and competing methods on 37 consensus viability profiles from the Achilles dataset. Methods: Random Forest (RF), Gaussian Process Regression (GPR), Multiple Pathway Learning (MPL), ensemble of MPL and Random Forest (MPL-RF), L2 regularized linear regression (GLM-dense), L1 regularized linear regression (GLM-sparse). B) Number of times an algorithm produced the best RMSE on a prediction task. To prevent small relative RMSE differences from having a biasing effect on the win counts, for each task we consider all algorithms with RMSE within 1% of the min RMSE to be joint winners. For that reason total win counts add up to more than the number of regression tasks. C) AKLIMATE's top 10 most predictive feature sets and top 50 most predictive features for the task of predicting MDM4 shRNA knockdown viability. Figure organized as Fig. 4.1. D) RMSEs of KRAS AKLIMATE models with and without the use of mutational profiles for 8 key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing. E) AKLIMATE's top 10 most predictive feature sets and top 50 most predictive features for the task of predicting KRAS shRNA knockdown viability when mutation features are used (feature and feature set weights averaged over 10 train/test splits). Figure organized as Fig. 4.1.	98
4.6	Method performance on predicting cell line viability after shRNA gene knockdowns measured by A) Spearman correlation. B) Number of times an algorithm produced the best Spearman correlation on a prediction task. C) Pearson correlation. D) Number of times an algorithm produced the best Pearson correlation on a prediction task.	99
4.7	RMSE scores for 37 shRNA knockdown prediction tasks. Rows correspond to individual gene knockdowns; columns represent different methods. To highlight differential performance within each task, rows have been centered by subtracting the median. Lower centered scores represent lower RMSE (better performance). . . .	100
4.8	AKLIMATE results highlighting the top 10 most informative feature sets and top 50 most informative features for the task of predicting FOXA1 shRNA knockdown viability. Figure organized as Fig. 4.1.	101

4.9	AKLIMATE results highlighting the top 10 most predictive feature sets and top 50 most predictive features for the task of predicting KRAS shRNA knockdown viability when no mutation features are used. Feature and feature set weights averaged over 10 train/test splits. Figure organized as Fig. 4.1.	102
4.10	Metrics for PIK3CA AKLIMATE models with and without the use of mutational profiles for 8 key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing.	103
4.11	Metrics for CTNNB1 AKLIMATE models with and without the use of mutational profiles for 8 key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing.	104
4.12	Data type importance proportions for 37 shRNA knockdown prediction tasks. Each column corresponds to the relative contributions (normalized to a sum of 1) of the different input data types for a particular prediction task. Relative contributions are computed by summing up the model-assigned feature importance scores within individual data types.	110
4.13	Comparison of AKLIMATE versus different ways of combining AKLIMATE component RFs. Ensemble superlearner component RFs - learning a regularized linear regression of predictions from component RFs; ensemble average component RFs - taking the average of the predictions of component RFs; top component RF - only using predictions from the top ranked RF. A) UCEC TCGA MSI prediction. B) METABRIC Breast Cancer survival prediction. C) Achilles shRNA knockdown prediction. P-values for Wilcoxon signed-rank test pairwise comparisons.	111
4.14	Overview of AKLIMATE. AKLIMATE takes as inputs multiple data types and a collection of feature set groups. AKLIMATE first trains RF learners, one for each feature set, with all available multi-omic features that map to the set in question. The RFs are then ranked by their OOB performance and the top K are converted to RF kernels. Finally, RF kernels are used as input in an elastic net MKL to produce the final predictions. Elastic net hyperparameters are learnt via cross-validation.	118

4.15	RF kernel construction. The RF Gram matrix is a geometric mean of the Hadamard product of three component similarity matrices, each capturing a different aspect of the RF model for a particular feature set: K1 - similarity over RF tree label predictions (in the case of classification, probability of belonging to a class), K2 - similarity over RF tree sample leaf indices, K3 - proportion of times samples are assigned to the same RF tree leaf.	124
5.1	Minimal Feature predictors of TCGA subtypes. TCGA subtypes are defined on the basis of tens of thousands of features, sometimes spanning several different platforms. However, the subtypes can often be characterized by a handful of covariates that can predict each subtype with a high degree of accuracy. For example, the four subtypes of gastric adenocarcinoma identified in the TCGA marker paper [107] can be faithfully recapitulated by two clinical variables (EBV+, diffuse histological subtype) and two composite features (MSI, chromosome instability). Figure panels taken from [107]. .	139
5.2	Subtype decision tree for the iCluster integrative clustering solution of the combined TCGA+CHOL cohort. The decision tree is derived from a hierarchical clustering dendrogram using iCluster latent variables as features. Numbers within each node state the size of the sample group available for that split. Earliest splitting nodes represent clusters that are most dissimilar from the rest, while later splits indicate less dissimilarity between groups. Each cluster is named after a clinical or molecular feature that it is highly enriched for. Figure created by Ronglai Shen and Esther Drill for the TCGA TMP AWG.	141
5.3	Performance of predictive models using A) 8 significantly enriched biomarker features and B) full AKLIMATE models. Models in A) were Random Forests grown to maximum tree depth, with 5000 trees and default mtry (rounded $\sqrt{8}$). Models in B) were AKLIMATE runs with the same hyperparameters, pathway collections and filtering steps described in Chapter 4. Barplots describe per-subtype balanced accuracies over 25 repetitions of 5 fold cross-validation (same folds in A) and B)). Balanced accuracy computed by taking the average of the sensitivity and specificity for each subtype.	142

5.4	Comparison of balanced accuracy for full and reduced AKLIMATE models. The full models are trained on 25 repeats of 5 fold cross-validation splits, as described in Fig. 5.3B. The reduced models are RFs trained on the top K most predictive features from the full AKLIMATE model, for different values of K , using the same folds as the full AKLIMATE model. Each reduced model is evaluated on the data fold that the full AKLIMATE model did not see during training. The grey interval around each line denotes one standard deviation from the mean (each mean is denoted by a dot).	145
5.5	A) AKLIMATE results highlighting the top 20 most informative feature sets and top 50 most informative features for predicting iCluster LIHC+CHOL subtypes. Feature and feature set weights are averaged over 25 repeats of 5 fold cross-validation. Figure organized as Fig. 4.1. B) Network view of the top 10 feature sets and the top 50 features. Feature sets encoded as squares, features as circles, with different colors for different data types. The size of each node is proportional to the normalized weight assigned by AKLIMATE to the respective feature/feature set. This is a bipartite graph, with features linked to feature sets they are a member of. To improve interpretability, each feature set has a uniquely colored cloud that maps out its member features.	148
5.6	Extended set of informative features within the WNT signaling pathway. Features shown are members of the pathway that also score as one of the top 500 most relevant features overall. Figure organized as Fig. 4.1.	148
5.7	Extended set of informative features within the TGF β signaling pathway. Features shown are members of the pathway that also score as one of the top 500 most relevant features overall. Figure organized as Fig. 4.1.	149

List of Tables

2.1	Combinatorial Analysis of significant MRs in CRPC.	42
3.1	Mean Spearman correlations of MPL, RF, and the 3 top performing methods from subchallenge 1 for high-confidence shRNA profiles. Individual correlation statistics for non-MPL DREAM9 methods downloaded from www.synapse.org (synapse ID syn2384331). . .	52
3.2	Top 15 canonical pathways enriched for genes with highly predictive MPL models.	56
3.3	Mean Spearman correlation for different (λ_1, λ_2) values	69
3.4	Mean Spearman correlation for different σ values	70
3.5	Mean Spearman correlation for different data type combinations. Both EXP and CN were filtered as described in the text.	70
3.6	Mean Spearman correlation for different gene set collections . . .	70
4.1	Most informative feature sets for breast cancer survival prediction in METABRIC data. Weights are averaged over 50 train/test splits. The table lists the top 20 most relevant feature sets, out of 1836 feature sets with a non-zero weight in at least one train/test split. Weights are normalized to sum to 1. Aliases for the top 10 feature sets are included in brackets - they provide a more descriptive representation of biological functionality, based on information gathered from the source publication. The aliases are also used in Fig. 4.4 and throughout the text.	96
5.1	Number of features used in the full AKLIMATE model, by data type.	143

List of Algorithms

1	AKLIMATE	131
---	--------------------	-----

Abstract

Prediction of cancer phenotypes through the integration of multi-omic data and
prior information

by

Vladislav J. Uzunangelov

High-throughput data have become ubiquitous in the study of biological phenomena. We can now query cellular state at higher resolution, giving us better insight into complex diseases. For example, there are currently tens of thousands of cancer patients with simultaneous copy number, mutation, methylation, mRNA, miRNA and protein level profiles. Furthermore, cellular perturbations are increasingly characterized on the multi-omic level. These experiments uncover important dependencies among genes, their products and environmental conditions - relationships that accumulate in a growing number of databases. However, the integration of such prior pathway knowledge with new heterogeneous genomic measurements in an interpretable model remains a formidable challenge that is still not fully solved.

My thesis presents three different approaches which incrementally address that problem. First, I present a feature engineering method (hVIPER) that infers kinase protein activity levels in a pathway-informed manner. Next, I develop one of the joint winners of the DREAM9 Gene Essentiality Prediction Challenge - a Multiple Kernel Learning algorithm with multi-omic pathway-derived kernel functions (MPL). Finally, I improve upon the DREAM9 winner by introducing empirical kernel functions computed through Random Forest tree ensembles (AKLIMATE). AKLIMATE outperforms state-of-the-art methods in diverse phenotype learning

tasks, including predicting microsatellite instability in endometrial and colorectal cancer, survival in breast cancer and shRNA knockdown response in CCLE cell lines. In conclusion, I briefly demonstrate how AKLIMATE can be adapted to the development of multi-omic minimum-feature predictors for patient subtypes.

To my parents Jordan and Antonia,
because you gave me all you could and more.

To my brother Alexander,
because you've always inspired me to keep going.

To my wife Terrie,
because you are here for me when I need it.

We walked this road together.

Acknowledgments

I'd like to thank my advisor Josh Stuart and my committee members Christopher Benz and David Draper for all their help and guidance over the years.

Special gratitude is reserved for the members of Stuart Lab, in particular Chris Wong for being main data organizer, main proofreader, main supporter and a friend, Evan Paull and Artem Sokolov for their advice and effort in winning DREAM9, and Rosalyn Huffman for making our lab run harmoniously.

Many compliments to Christina Yau for being a great collaborator and mentor in TCGA projects, and to Joshi Alunkal and the West Coast Dream Team for our prostate cancer work.

Finally, a big thank you to Julian and Flora, as well as all folks at the Pacific Edge Climbing Gym for keeping me sane over the years.

Chapter 1

Introduction

Modern cell biology is a science of data integration. Experiments are now routinely set up to collect genomic, transcriptomic and proteomic measurements, often in the same patient cohort. The Cancer Genome Atlas (TCGA) is one such example - the consortium generated a compendium of over 10,000 samples that has been richly characterized across multiple data modalities[66]. By examining these data types both individually [8][24][66] and integratively [66], TCGA has led to major breakthroughs in the molecular-level understanding of cancer initiation and progression.

A great majority of published biological experiments, ranging in size from TCGA-level analysis to small-cohort expression studies, produce signatures of differential features that capture contrasts between subgroups of the examined cohorts (e.g. [42]). Such signatures can serve two purposes - on one hand, they inform our growing understanding of gene-gene interaction networks and the genetic, signaling and metabolic pathways they describe. On the other hand, they can be directly used in unsupervised [148] or supervised [59] learning tasks as

a way to transfer prior knowledge to current analysis. In fact, many of the Dialogue on Reverse-Engineering Assessment and Methods (DREAM)[137] challenges have demonstrated that the incorporation of pathway or signature (PS) information in predictive modeling can boost performance across a wide range of tasks [32][147][59]. As a consequence, efforts to catalogue and update PS information have led to the proliferation of (often heavily overlapping) PS collections and databases, with sizes varying from a few dozen to 10,000's. Their focus ranges from the most narrow view of "canonical" pathways [76][46] to ones that specialize in a particular data type [55][73][68] to general compendiums that aggregate many disparate sources [21][93][114].

The emergence of next generation data-generating technologies in the last decade has clearly fueled the rapid recent growth of PS collections - more multimodal data profiles introduce more novel signatures. However, there is still paucity of methods that can truly take advantage of 10,000's of heterogeneous data features and 10,000's of potentially relevant PS feature sets. My thesis provides a sequence of progressively better solutions to that problem, leading up to the introduction of a novel machine learning method that natively integrates all sources of information at scale.

In Chapter 2, I lay the foundations for multi-modal data analysis by describing a feature engineering method (hVIPER) that solves one aspect of a problem critical to many experimental designs - what if we can't generate the data that is most salient to our investigation? In particular, hVIPER approximates kinase protein state by an activity score computed from expression data. hVIPER uses a pathway-informed hierarchical application of an existing method for the inferral of transcription factor activities [3]. Chapter 2 gives examples of A) the integra-

tion of hVIPER activities with other data types for a deeper characterization of cancer subtypes and B) hVIPER utility in identifying novel therapeutic targets in treatment-refractory castration-resistant prostate cancer. Furthermore, hVIPER activities are one of several data modalities used for the prediction of shRNA knockdown profiles in Chapter 4.

Chapter 3 transitions to the creation of learning models that can handle both heterogeneous data and PS prior knowledge. One commonly used approach for multimodal data integration is Multiple Kernel Learning (MKL)[7] - a framework for the discovery of optimal (generally conical or convex) combinations of pairwise sample similarities (kernel matrices). In an MKL setup, different data types can be encoded as distinct kernel matrices by utilizing data-type specific linear or non-linear transformations. Chapter 3 introduces Multiple Pathway Learning (MPL)-an MKL instance in which each kernel serves the dual purpose of encoding a specific data type as well as a particular PS gene set. As a consequence, an MPL solution is in effect the most predictive meta-pathway we can construct with available PS gene sets as building blocks. MPL co-won the DREAM9 gene essentiality challenge[59] - chapter 3 describes how we achieved that milestone and what insights we gained from the different predictive models.

Despite its obvious advantages and excellent predictive performance, MPL is not without shortcomings. In particular, it cannot assign predictive importance to features in the original input space and it cannot model inter-data type feature dependencies. In chapter 4, I introduce AKLIMATE - an algorithm that resolves MPL’s deficiencies in feature interpretation and integration. It still operates with PS-based kernels in an MKL framework, but each kernel matrix is derived through the clever interpretation of supervised decision tree ensembles. This novel manner

of kernel construction is in large part responsible for AKLIMATE’s superior performance when compared to standard MKL methods of various flavors. Finally, Chapter 5 presents ideas for further AKLIMATE improvements or applications.

This dissertation introduces a modern, flexible and extensible framework for the native integration of multimodal data and prior knowledge. AKLIMATE’s unique blend of Random Forests, MKL and stacked learning provides a blueprint for the creation of interpretable, high-performing and scalable ensemble models. As this body of work demonstrates, such models can make crucial contributions to a wide range of genomic prediction tasks.

Chapter 2

Hierarchical VIPER (hVIPER)

In this chapter I describe a feature engineering technique (hVIPER) that uses pathway information to infer transcription factors (TF) and kinase protein activity. The technique extends a previous approach called Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) [3] that derives TF activities from the expression of their targets. hVIPER modifies the original VIPER model in two important ways. On one hand, hVIPER uses a hierarchical inference approach to compute kinase activities (in addition to TF ones). Characterizing the kinome behavior is crucial in translational medicine (e.g. drug development and drug therapy) as kinases are generally much easier to target pharmacologically than TFs. On the other hand, hVIPER activities are derived from pre-assembled compendiums of TF and kinase targets instead of the empirically reconstructed regulator-target networks used in VIPER. The use of literature-based target sets allows for the intuitive incorporation of past experimental results into new data analysis. Furthermore, they are generally more accurate than de novo data-inferred TF regulomes, at the expense of reduced TF coverage [52].

I present two applications of hVIPER to the analysis of dysregulated networks in A) uveal melanoma (UM) and B) metastatic castration-resistant prostate cancer (CRPC). I was responsible for all hVIPER results in the UM marker paper[117]. Furthermore, in collaboration with Christina Yau (provider of PARADIGM[160] analysis), I contributed the network characterization of the main UM subtypes which required integrating hVIPER activities with lncRNA and PARADIGM features. The introduction, data generation, and subtype analysis included here are attributable to all authors of [117].

The CRPC study has been submitted for publication. It is a collaboration under the umbrella of the Stand Up 2 Cancer West Coast Dream Team (WCDDT), with main contributions from Josh Stuart’s lab at UCSC and Joshi Alumkal’s lab at OHSU. I contributed the AR signature and hVIPER analyses. The regulatory network discovered by hVIPER led to the main finding of the manuscript - a CREB1-influenced dysregulation circuit that could lead to potential novel targeted therapies for CRPC patients. The introduction and data generation are attributable to all authors, with wet lab validation contributed by Joshi Alumkal and his group.

2.1 Uveal Melanoma

2.1.1 Introduction

Uveal melanoma (UM), which arises from melanocytes resident in the uveal tract, is the second most common melanoma subtype after cutaneous melanoma (CM) [134][163]. Although both UM and CM tend to occur in people with light iris color and fair skin [167], their clinical and molecular characteristics are very

different [33][170]. While primary UM is treated with either surgery or radiation and has a low local recurrence rate, up to 50% of UM patients develop distant metastatic disease, often to the liver, after treatment of the primary tumor. At present there are no effective therapies for metastatic UM, and most patients survive less than 12 months after diagnosis of metastases [45][22]. UM displays chromosome aberrations and gene mutations that correlate strongly with clinical outcome and are not present in CM. Loss of one copy of chromosome 3 (monosomy 3, M3) in UM is associated with an increased risk of metastasis and a poor prognosis [39][133]. Loss-of-function mutations in BAP1, which is located on 3p21, have been identified in M3-UM [61], and decreased BAP1 mRNA and protein expression, indicating BAP1 aberrancy, are highly correlated with the development of UM metastases [75][84]. Currently either disomy 3 (D3) versus M3 status or a 12-gene microarray-based gene expression panel is used to determine whether a patient is in a low- or a high-risk prognostic group [60][152]. Recent analysis of a large D3-UM cohort showed SF3B1 mutation to be associated with an intermediate risk of developing later-onset metastatic UM [173]. Despite prognosis being clearly correlated with the expression of a small panel of marker genes, with M3, and with BAP1 aberrancy or SF3B1 mutation, the molecular pathways involved in the development of metastatic disease have not been elucidated. In this Rare Tumor Project of The Cancer Genome Atlas (TCGA), we performed a global and integrated molecular characterization of 80 primary UM, seeking to generate insights into biological processes that underlie UM tumors that have distinctly different prognoses.

2.1.2 Results

We used RNA-seq data to profile the expression of 20,531 mRNAs and of 8,167 long non-coding RNAs (lncRNAs) and processed transcripts, and identified four-cluster consensus solutions for both mRNA and lncRNA (Fig.2.1A&B). The two clustering solutions exhibited very high concordance. This concordance was maintained in subtypes generated from other data types as well - very high for PARADIGM-based clusters and reasonably strong for miRNA-, methylation-, and copy number-based ones (Fig.2.1B). Unless specifically stated, the rest of the chapter describes the lncRNA-based solution. D3-UM divided into transcription-based clusters 1 and 2, M3-UM into clusters 3 and 4, and the 12-gene panel's two prognostic groups were each further separated into two groups.

As expected, M3-UM patients had a significantly worse prognosis than D3-UM (Fig.2.1C). BAP1 alterations are widely prevalent in poor-prognosis clusters 3 and 4 - such alterations lead to biallelic BAP1 loss since patients in these subtypes are already characterized by the loss of chromosome 3. Furthermore, M3/BAP1-aberrant tumors from both poor-prognosis clusters map to the same DNA methylation cluster (Fig.2.1B), suggesting the existence of a global M3/BAP1-aberrant methylation pattern.

Interestingly, even though the poor-prognosis clusters group together based on survival, large scale copy number events, and the 12-gene prognostic signature, they have very different transcriptomic profiles, suggesting the operation of distinct biological processes. A significant difference between cluster 3 and cluster 4 is that the latter contains all samples with high inferred CD8 T cell infiltration levels (Fig.2.1B). The association between immune infiltration and M3/poor prognosis has been documented previously [17]. Leukocyte infiltration alone, however,

is not sufficient to explain cluster 4 as only $\sim 60\%$ of the tumors in it exhibit high leukocyte fraction.

To further characterize the differences between the two poor-surviving clusters, we performed PARADIGM and hVIPER (see 2.1.4.2) analysis. Each of the methods provides a measure of pathway activity - PARADIGM by integrating expression and copy number on a network of gene interactions and hVIPER by approximating master regulator (MR) activity from the expression or inferred activity of an MR's targets. Since the two methods provide distinct pathway-informed views of the patients, we hypothesized that combining the results of the two approaches will give us the most complete description of the different biological processes underpinning poor-survival phenotypes.

On the feature level, PARADIGM and hVIPER differential features tend to show only moderate overlap (Fig.2.2, also see Section 2.1.4.3). However, the overlap is much stronger on the pathway level, with the two methods providing complementary support for the dysregulation of key biological processes.

In cluster 3, PARADIGM identified higher activities of key transcription factors FOXA1 and FOXM1, as well as elevated levels of MAPK1 and AKT, indicating high cellular cycling and cell proliferation (Fig.2.2A and Fig. 2.3A). Those findings were supported by the elevated hVIPER scores of E2F1 (proliferation) and MAPK1 (Fig.2.2B and Fig. 2.3A). Furthermore, DNA Damage Response (DDR) appears to be downregulated, with support from both PARADIGM (lower activity of the DNA Damage Process abstract) and hVIPER (lower activity of ATR and CHEK2). Finally, the downregulation of hypoxia response can be inferred from the lower activities of the HIF1 α /ARNT complex (PARADIGM) and the HIF1 α regulator feature (hVIPER).

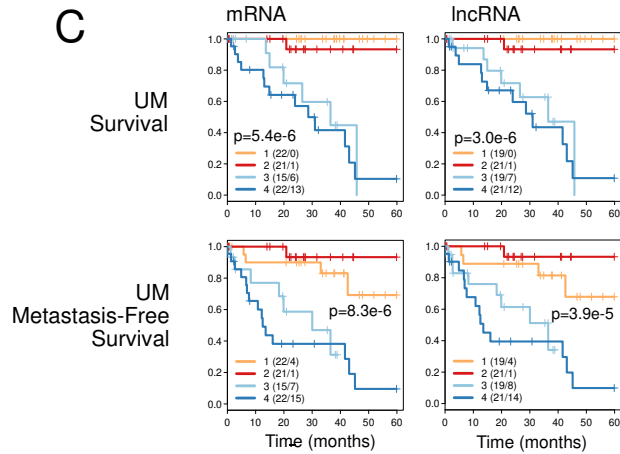
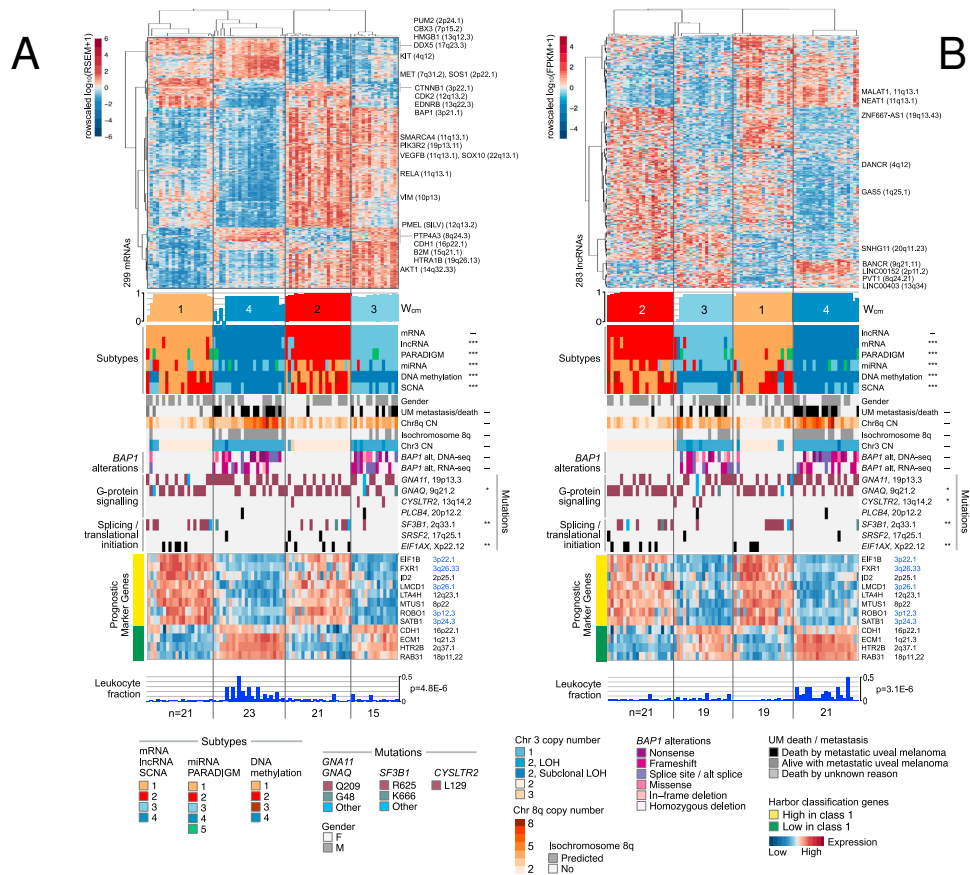


Figure 2.1: Clustering solutions based on A) mRNA and B) lncRNA features. Covariate annotation tracks show selected genomic and clinical features. The lower heatmap displays the expression profiles of 12 genes used in a prognostic test for the risk of developing metastasis [60], with blue text highlighting genes that are on chromosome 3. High-risk primary tumors show low expression of eight of these genes and high expression of four genes (yellow versus green panels at the left). BAP1 structural alterations that include alternative splicing and rearrangements were detected by assembly of RNA-seq and DNA-seq data. Leukocyte fraction was estimated from DNA methylation data. LOH, loss of heterozygosity. *, **, *** p value < 0.1, 0.01, 10⁻³, Fisher’s Exact or Chi-square test. C) Kaplan-Meier curves for overall and metastasis-free survival for mRNA and lncRNA subtypes, with log-rank p-values. Figure panels taken from [117].

In cluster 4, both PARADIGM and hVIPER capture strong immune-related activity, concordant with the high leukocyte infiltration levels previously discussed. In particular, the JUN/FOS complex (identified by both methods) and JAK2-related signaling (IL23/JAK2, IL27/JAK2 complexes and JAK2-STAT1/3 signaling, all supported by both methods) were highly upregulated (Fig. 2.2 and Fig. 2.4A). Unlike cluster 3, increased activity of RAD21, ATM and ATF2 (hVIPER) as well as the DNA Damage Process abstract (PARADIGM) pointed towards upregulation of DDR. In another cluster 3 contrast, strong MYC signaling (hVIPER), possibly through the MYC/MAX/MIZ1 complex (PARADIGM), coupled with higher levels of HIF1 α (hVIPER, HIF1 α /ARNT complex in PARADIGM) suggest upregulated hypoxia response.

The difference in DDR activity between the two poor-surviving clusters is particularly interesting in light of the fact that BAP1 has been implicated in the regulation of DNA damage response/repair - BAP1 loss of function leads to inefficient DDR [71][174]. As the latter is similarly prevalent in clusters 3 and 4, a fruitful future topic of investigation would be the elucidation of other contributing

factors for the divergent DDR response in M3 patients.

On a similar note, and related to divergent hypoxia signaling, PARADIGM and hVIPER could only produce a partially consistent view of MYC activity. MYC is an oncogenic transcription factor located on chromosome 8 (8q24.21) whose involvement in UM progression has been suggested previously [102][157]. As M3 and 8q gain show a high correlation (Fig. 2.1), we expected MYC signaling to play a prominent role in the networks describing poor-surviving patients. Interestingly, both PARADIGM and hVIPER produced some unexpected results - PARADIGM's MYC/MAX complex showed higher activity in cluster 4, but not in cluster 3 (Fig. 2.2A). Furthermore, it had unexpectedly high activity in cluster 1, a group of good survivors with D3/8q-normal tumors. The MYC/MAX/MIZ1 complex showed similarly strong signal in cluster 4, much more neutral signal in cluster 1, and decreased activity in the other group of M3 poor survivors (cluster 3). Similarly, hVIPER registered increased MYC activity in cluster 4 (but even more so in the good-survivors of cluster 1), yet MYC downregulation in cluster 3 (Fig. 2.2B). MYC's partner in the MYC/MAX complex was highly active in M3/8q-gain cluster 3, but also active in the D3 good survivors of cluster 2 (although the latter exhibits a pattern of 8q gain as well).

Taken together, the evidence points toward increased MYC activity in cluster 4 (Fig. 2.4A), but is a lot more inconclusive for the other lncRNA clusters (although suggesting downregulation in cluster 3 - Fig. 2.3A). One possible explanation for this lack of clarity is that MYC can either upregulate or repress its gene targets, depending on the activity of different complexes it is a part of [85]. Another explanation is that the interplay between MYC and its regulators is causing divergent MYC activity. For example, PVT1 is a lncRNA situated on

chromosome 8 (8q29.21) that is one of the most differentially expressed transcripts in poor-prognosis lncRNA clusters 3 & 4. Furthermore, its locus is adjacent to the MYC locus and it is coamplified with MYC in 8q-gain UM tumors. PVT1 has been previously shown to stabilize MYC protein levels [28] and enhance its activity [153]. A more precise examination of the MYC-PVT1 interplay could be crucial in explaining MYC's behavior in the context of UM.

As showcased by PVT1, lncRNAs can play significant roles in cancer development and progression. The interactions between lncRNAs and key regulators, however, are often poorly understood. As a first step towards constructing a lncRNA-MR interactome, I constructed cluster-specific networks from the most differential findings of hVIPER, PARADIGM, and a separate lncRNA differential analysis (see 2.1.4.3). The recovered network for cluster 4 has 709 correlations between 26 lncRNAs, 70 hVIPER and 29 PARADIGM features (Fig.2.4A). LINC00152 and BANCR are among the most highly connected and the most differentially expressed lncRNAs in the cluster 4 network (Fig.2.4A, Fig.2.1B). The expression levels of these lncRNAs are strongly correlated with regulator activities in previously discussed key pathway processes - e.g. immune-related JAK2, JUN/FOS and STAT complexes, mediators of DDR, and MYC/MAX & HIF1 α hypoxia-related regulators (Fig.2.4B).

Similarly, the recovered network for cluster 3 contains 188 correlations between 10 lncRNAs with 21 hVIPER and 24 PARADIGM pathway features (Fig.2.3A). In addition to being much smaller than the cluster 4 network, it only contains lncRNAs that are poorly characterized (or not characterized at all) in cancer settings. Nevertheless, most of these lncRNAs exhibited high correlation with activated transcriptional regulators of proliferation (FOXM1, E2F1, PRKACA)

and negative correlation with low DDR pathway activity (ATR, CHEK2, CDK1, DNA Damage Process abstract, Fig.2.3B).

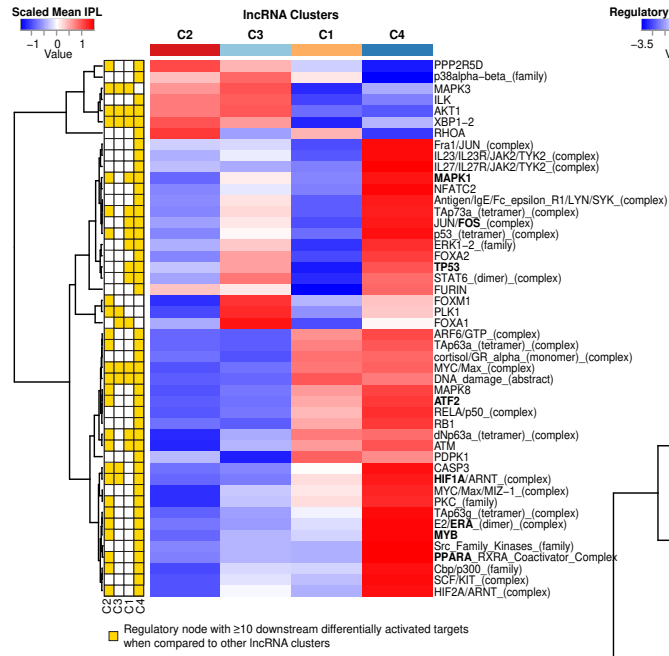
2.1.3 Data

Eighty primary UM tumors were available for multiplatform analysis. Cancer cell contents were high based on ABSOLUTE [20] (median purity = 0.95), DNA methylation-derived leukocyte fraction, and histopathological assessment. All cases were $\geq$ T2 (seventh edition of the AJCC TNM-staging system); $\sim$ 10% of patients developed another primary malignancy.

2.1.3.1 PARADIGM Superpathway

Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases (<http://pid.nci.nih.gov>) and the Reactome database (<http://reactome.org>). Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee’s HUGO symbols using mappings provided by HGNC (<http://www.genenames.org>). Altogether, 1,524 pathways were obtained. Interactions from all of these sources were combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. “cell cycle” and “apoptosis”) were retained and are collectively referred to as pathway “features”. The resulting pathway structure contained a total of 19,504 features, representing 7,369 proteins, 9,354 complexes, 2,092 families, 82 RNAs, 15 miRNAs and 592 abstract processes.

A



B

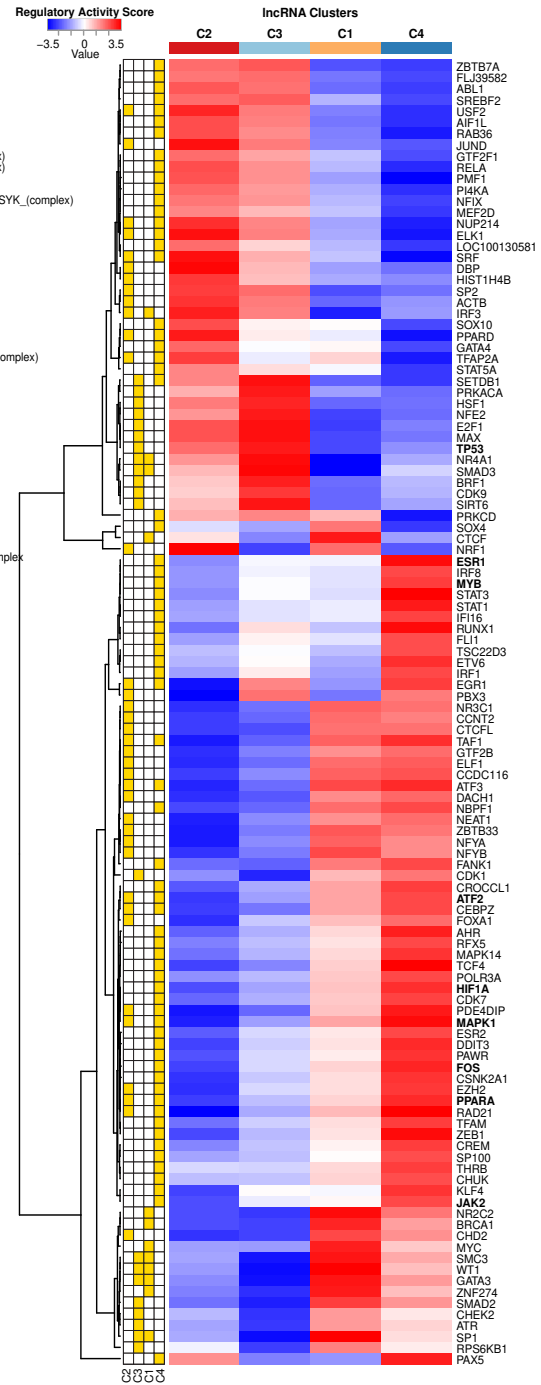


Figure 2.2: Significant findings from the A) PARADIGM and B) hVIPER pathway biomarker analysis (see 2.1.4.3). A) The heatmap displays scaled mean IPLs of regulatory hubs that have at least 10 downstream targets and are differentially active in at least one of the one-vs-all-other cluster comparisons. The panel at the left identifies the clusters in which a hub is differentially active (yellow). B) The heatmap displays master regulators (MRs) that have differential activity levels in at least one of the one-vs-all-other cluster comparisons. The panel at the left identify the lncRNA cluster in which MRs are differentially active (yellow). For each cluster, gene biomarkers with similar findings across the two methods are highlighted in boldface. The similarity can be based on: 1. complexes or families with components identified by both methods, 2. genes within the same pathway showing complementary inferred activity, or 3. abstract processes linked to any of the above. Figure panels taken from [117].

2.1.3.2 hVIPER curated Transcription Factor and Kinase Regulome

A compendium of TFs and their targets (TF regulons) were created by combining information from four databases:

1. SuperPathway [128]: This is the same interaction network described in 2.1.3.1. Only links that correspond to regulation at the transcriptional level were retained for VIPER and hVIPER use.
2. Literome [112]: The network was filtered to include only transcription links in which the regulator is a known TF.
3. Multinet [78]: The network was reduced to links that correspond to regulation on transcriptional level.
4. ChEA [86]: Data from the Gene Expression Atlas [110] was used to filter the inferred links in the ChEA database. Specifically, the context likelihood of relatedness (CLR) method [47] was used to compute a measure of association between every pair of genes. The top 10% of gene pairs based on the CLR

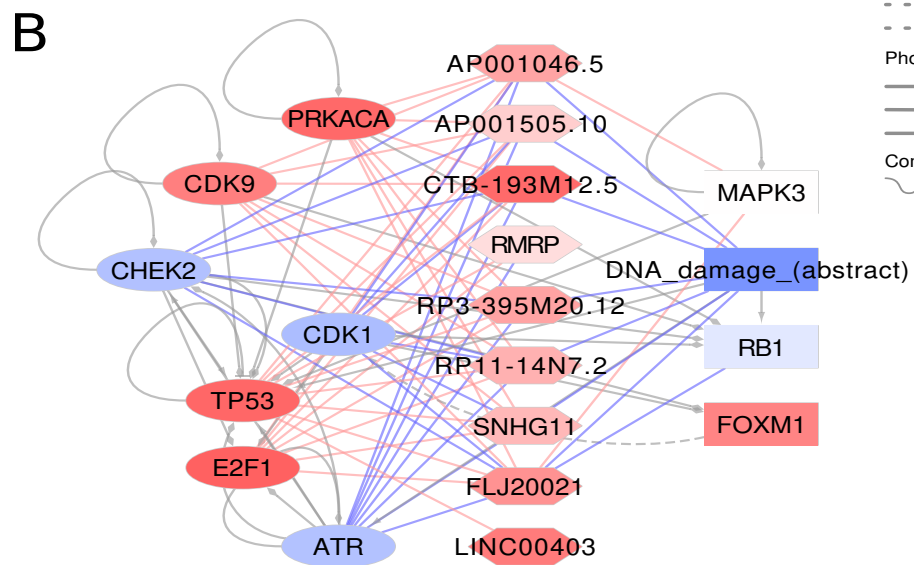
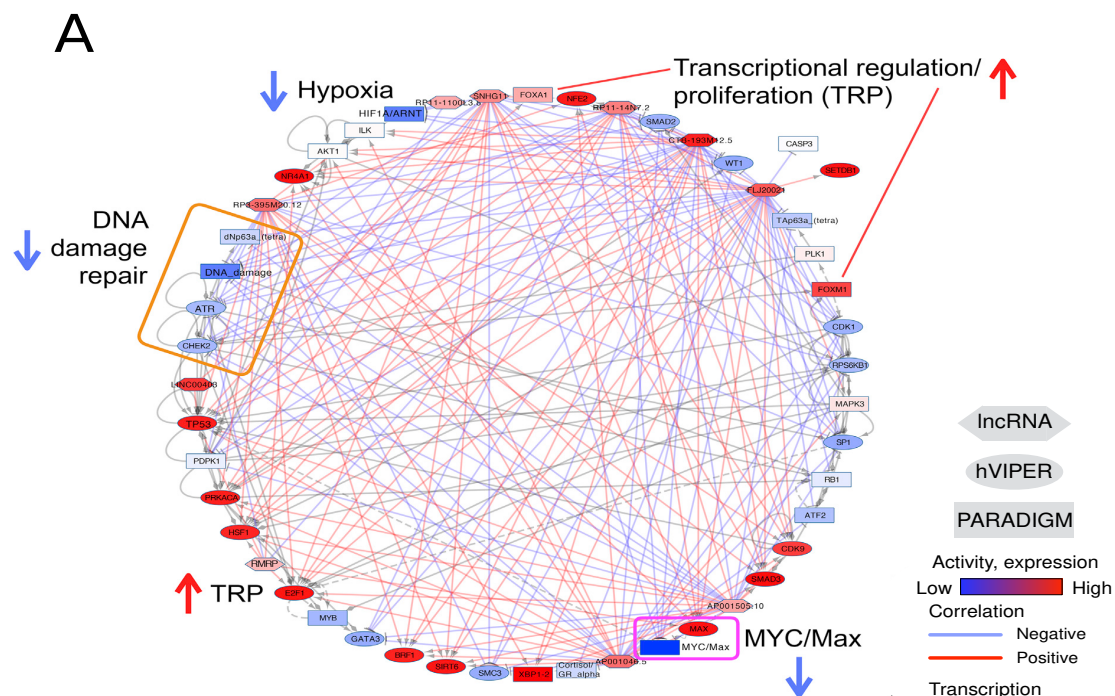


Figure 2.3: A) Correlation network for lncRNA clusters 3, showing PARADIGM pathway features, hVIPER regulators, and lncRNAs. Red and blue edges indicate Spearman correlations ($|\rho| > 0.5$) between the expression of a differentially expressed lncRNA and the inferred activity of a differentially active PARADIGM or hVIPER feature. The color of each node reflects differential expression for a lncRNA, and relative activity for a PARADIGM/hVIPER feature (red for over-expressed/active, blue for underexpressed/inactive).

B) A subnetwork highlighting the lncRNAs with high degree of correlation to regulators in the proliferation and DDR pathways.

Both panels generated by myself, with input from Christina Yau and Gordon Robertson, for the TCGA UM working group. Panel A) also appears in [117].

score were intersected with the ChEA network and the overlapping pairs were added to the final combined network.

The combined network includes 72,915 transcriptional regulatory links between 6,735 regulators and their targets. Only regulators with at least 15 targets were considered in the final analysis, which resulted in a final network consisting of 419 TFs with 58,363 total targets (covering a set of 12,754 unique targets).

Proteins identified as kinases in Manning [96] or Uniprot [30] were aggregated into a list of 546 kinases. Protein substrates were extracted from PhosphositePlus [68]. Kinase-substrate interactions were retained if the kinase appeared in the Manning-Uniprot kinase list and the kinase was identified as a human protein in the PhosphositePlus database. The final compendium consists of 5,388 links between 342 kinases and 2,290 unique substrates.

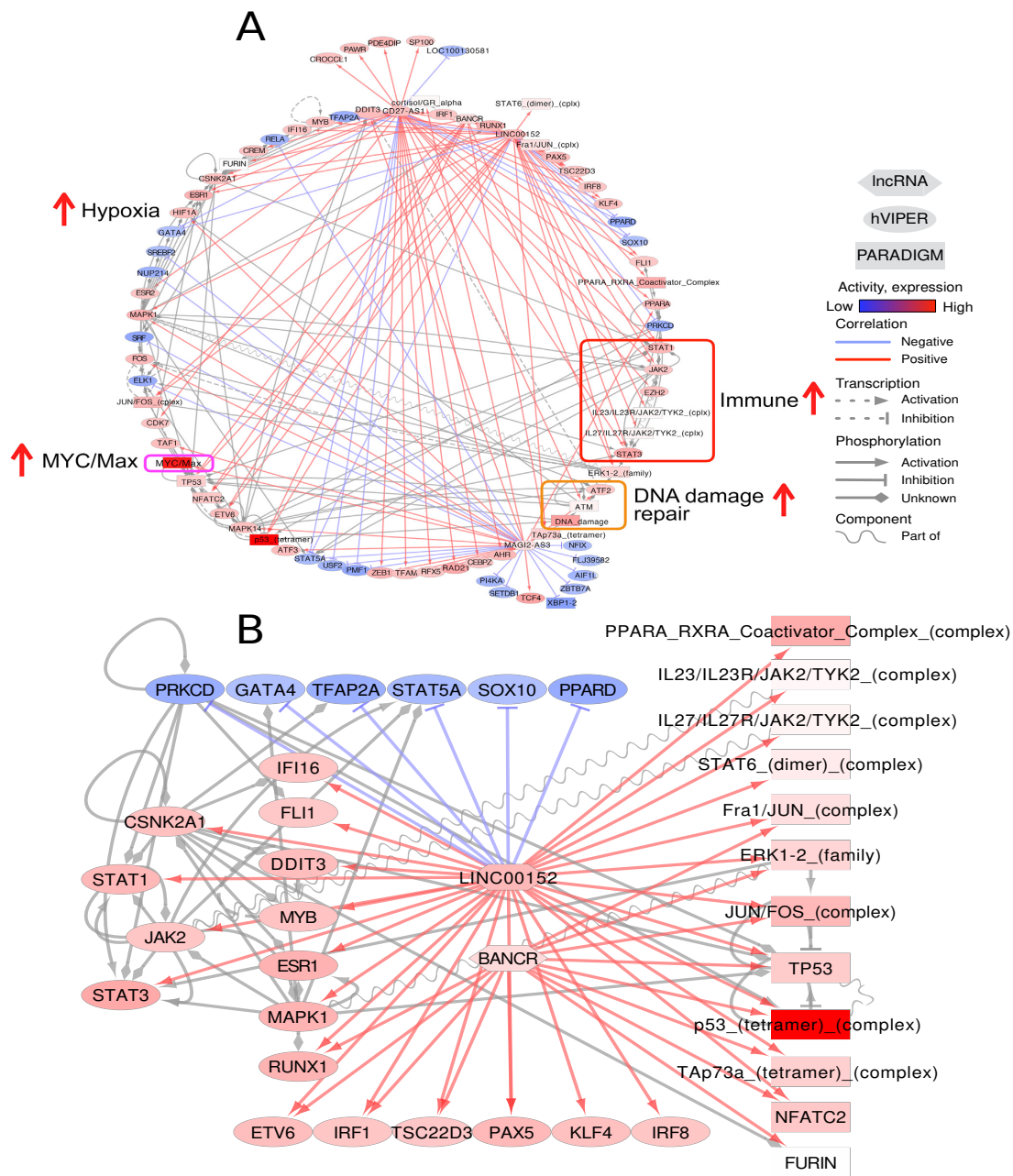


Figure 2.4: A) Correlation network for lncRNA clusters 4, showing PARADIGM pathway features, hVIPER regulators, and lncRNAs. Red and blue edges indicate Spearman correlations ($|\rho| > 0.5$) between the expression of a differentially expressed lncRNA and inferred activity of a differentially active PARADIGM or hVIPER feature. The color of each node reflects differential expression for a lncRNA, and relative activity for a PARADIGM/hVIPER feature (red for over-expressed/active, blue for underexpressed/inactive).

B) A subnetwork highlighting the breadth of associations of LINC00152 and BANC1 with regulators in the immune response, DDR, MYC/MAX and hypoxia signaling pathways.

Both panels generated by myself, with input from Christina Yau and Gordon Robertson, for the TCGA UM working group. Panel A) also appears in [117].

2.1.4 Methods

2.1.4.1 VIPER

VIPER scores predict the activity of a TF, either differentially between two cohorts, or within an individual sample (single-sample VIPER, ssVIPER). The activity score is derived from a combined view of the expression levels of each TF's transcriptional targets (the TF regulome). Traditionally the TF regulome is inferred from data using network reconstruction methods such as ARACNe [100]. In our approach we eschew the empirically derived interaction network for one constructed from literature-extracted "prior knowledge". Once a TF regulome has been selected, VIPER scores are computed as follows (for more details, see [3]):

1. Each gene in the TF regulome is assigned a particular mode of operation (MOR). The MOR can be one of the following relationships:
 - (a) Regulator clearly represses target.
 - (b) Regulator clearly activates target.

(c) Regulator regulates target non-monotonically.

The distributions of the three types of relationships are estimated from the overall density of all regulator-target correlations by fitting a mixture model of 3 Gaussian distributions. Each target gene then gets assigned a triplet of weights corresponding to its probability of belonging to each MOR class.

2. For multi-sample VIPER, a t-statistic derived from the difference in expression between the two classes of interest is computed for each gene. All genes are ranked based on their t-statistics to produce a gene signature. For ssVIPER, the ranked gene signature is based on the gene expression levels within the sample instead.
3. An enrichment score is calculated by ranking regulome members by the absolute value of the t-statistic/gene expression and computing the mean of the quantile transformed rank positions, weighted by $(1 - |MOR|)$ (one-tail approach).
4. A second enrichment score is calculated that uses the same t-statistic/gene expression ranks, but with ranks of MOR-repressed targets inverted. The enrichment score is again the mean of the quantile transformed rank positions, but it is now weighted by $|MOR|$ (two-tail approach).
5. The final score is the sum of the one-tail and two-tail enrichment scores.

A TF whose MOR-upregulated and MOR-downregulated target sets show consistent enrichment (i.e. the activated set is enriched for highly ranked genes and the inhibited set is enriched for lowly ranked ones, or vice versa) receives the highest (lowest) activity score.

2.1.4.2 hVIPER

VIPER/ssVIPER are well suited for the analysis of TF activity, because TF proteins are directly involved in changing the expression of their targets. Kinases, on the other hand, regulate their targets post-translationally. Since the expression levels of genes are often poorly correlated with the activity of the proteins they encode, mRNA represents a poor proxy to protein phosphorylation data. In the absence of the latter, the differential activity of a kinase can be estimated using a hierarchical approach (Fig. 2.5) in which activities are computed at two successive levels:

1. Level 1 activities are inferred for any regulator (TF or kinase) using single-sample VIPER (ssVIPER). ssVIPER infers these activities based on the expression of the regulator’s targets within individual samples. Note that the kinase activity score from level 1 analysis is interpreted as an inference about whether kinase targets are “poised” to be regulated, assuming that increased protein levels would often require an increase in mRNA production as a prerequisite.
2. Level 2 activities for kinases are inferred by performing a VIPER analysis on the level 1 activities computed in the previous step. For level 2, the targets of each kinase are restricted to those members that are themselves kinases or TFs (i.e. those proteins with level 1 imputed activities).

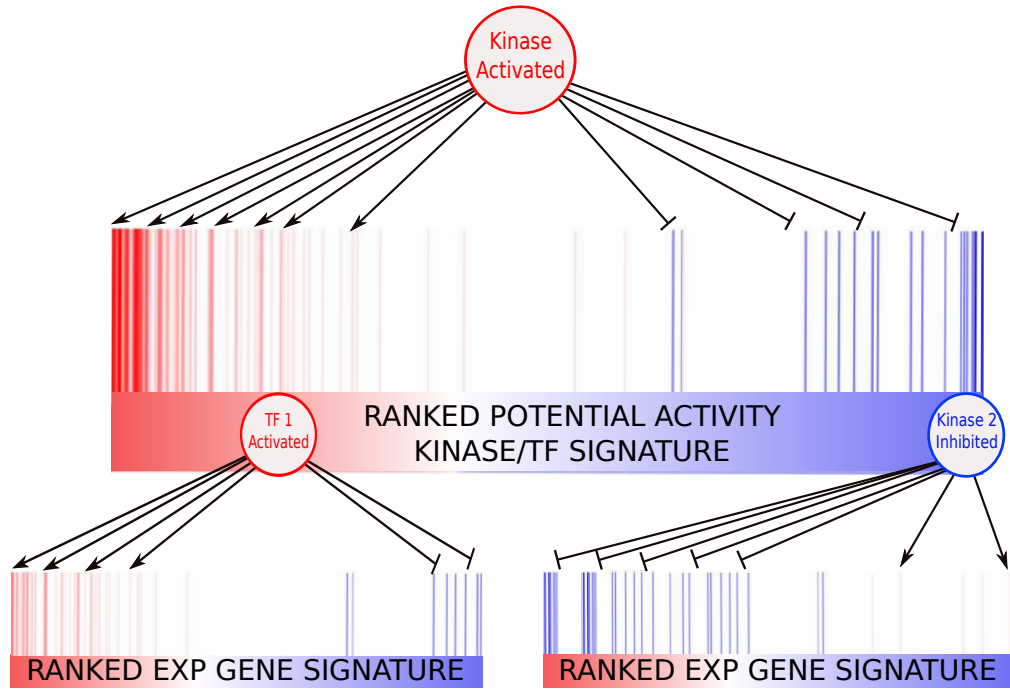


Figure 2.5: Hierarchical VIPER for the imputation of kinase activities. ssVIPER is performed for all TFs and kinases with sufficient number of targets in their regulons (bottom level). The inferred activities are used to rank regulators based on difference in mean activity across the dichotomy of interest. Standard VIPER analysis is then run with the kinase regulome and the regulator ranking to produce the final estimates for kinase activities (top level).

2.1.4.3 PARADIGM and hVIPER differential analysis of UM lncRNA clusters

PARADIGM

mRNA expression, SCNA, and pathway interaction data for 80 UM samples were integrated using the PARADIGM software [160]. Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes, using pathway interactions, and genomic and functional genomic data from each patient sample. Normalized gene-level RSEM RNA-seq expression data and thresholded SCNA data were obtained from Firebrowse (<http://firebrowse.org/>). One was added to all expression values, which were then log2 transformed and median-centered across samples for each gene. The log2 transformed, median-centered mRNA data were rank-transformed based on the global ranking across all samples and all genes and discretized (+1 for values with ranks in the highest tertile, -1 for values with ranks in the lowest tertile, and 0 otherwise) prior to PARADIGM analysis.

The PARADIGM algorithm infers an IPL for each feature that reflects the log likelihood of the probability that it is activated. PARADIGM IPLs of the 19,504 features within the SuperPathway (see 2.1.3.1) are available on Synapse (syn4556715, <https://www.synapse.org>). An initial minimum variation filter (at least 1 sample with absolute activity > 0.05) was applied, resulting in 15,502 concepts (5,898 proteins, 7,307 complexes, 1,916 families, 12 mRNAs, 15 miRNAs and 354 abstract processes) with relative activities showing distinguishable variation across tumors (syn4556729) for use in our differential pathway regulator analysis.

Differential pathway regulators of each lncRNA cluster were identified using the t-test and Wilcoxon Rank Sum test with BH FDR correction in a one cluster vs. all others comparison. Only features deemed significant (FDR $p < 0.05$) by both tests and showing an absolute difference in group means > 0.05 were selected. Interconnectivity between these pathway regulators within the PARADIGM SuperPathway was assessed, and regulatory hubs with ≥ 10 differentially activated downstream targets were selected. There were a total of 49 PARADIGM differential pathway regulators identified across the four lncRNA clusters. The mean IPL of the selected regulatory hubs was computed within each cluster and scaled across clusters to a mean of 0 and a standard deviation of 1. The resulting scaled mean IPLs are shown in Fig.2.2A.

hVIPER

The lncRNA clusters were dichotomized into one-vs-rest binary comparisons. For each comparison, VIPER/ssVIPER was run via the VIPER R package (<http://www.bioconductor.org/packages/release/bioc/html/viper.html>); and hVIPER was performed by extending the functionality of the package. Level 3 mRNA data and the curated TF and kinase regulomes (see 2.1.3.2) were used as inputs. Analysis was limited to TFs with at least 15 targets present in the expression data. Because the kinase regulome is much smaller than the TF regulome, cutoffs for minimum number of kinase substrates were reduced to 10 in the Level 1 analysis and 5 in the Level 2 one. All other settings were identical to those used for inferring TF activity.

Background models were computed by generating 1000 label permutations. Significance was evaluated by computing p-values against the background distri-

bution and applying a BH FDR correction. The final results provided activity estimates for 393 TFs and 62 kinases in each dichotomy of interest. VIPER features (TFs) with an $\text{FDR} \leq 0.10$ were retained. Since the kinase regulome is significantly sparser than the TF one, the FDR cutoff for hVIPER features was relaxed to 0.15. A total of 113 hVIPER differential pathway regulators were identified across the four lncRNA clusters. The differential activity for each of these regulators in each lncRNA cluster is shown in Fig. 2.2B.

lncRNA Pathway Regulator Correlation Networks

The FPKM expression of every lncRNA was correlated with PARADIGM per-sample IPL levels, and with the TF and kinase activities produced by hVIPER, using per-sample ssVIPER activity scores. For each lncRNA cluster, correlations between differentially active regulators and lncRNAs were retained if all four of the following criteria were satisfied:

1. The TF/kinase was identified as a differentially active pathway feature by PARADIGM or hVIPER for that cluster, as described above.
2. The lncRNA had a mean FPKM ≥ 5 .
3. The lncRNA had a SAM [155] multiclass FDR q-value ≤ 0.05 and the absolute value of its SAM contrast for the cluster was the largest compared to the absolute contrast values for all other clusters.
4. The absolute value of the Spearman correlation coefficient between the lncRNA and the regulator in question was ≥ 0.5 .

The filtered lncRNA-pathway regulator network for lncRNA Cluster 3 contains 188 correlations between 10 lncRNAs with 21 hVIPER and 24 PARADIGM pathway features. Similarly, the filtered lncRNA-pathway regulator network for lncRNA Cluster 4 contains 709 correlations between 26 lncRNAs with 70 hVIPER and 29 PARADIGM features. Figures 2.3A and 2.4A each shows the full network of selected regulators and their associated lncRNAs. For the full list of links, see https://tcga-data.nci.nih.gov/docs/publications/uvm_2016.

Both Figures 2.3 and 2.4 are augmented by protein-protein interaction and transcriptional regulation links extracted from PhosphositePlus and the Super-Pathway (see 2.1.3.2). In addition, regulators that were identified as consistent pathway features by both methods were displayed using the shape of the method that showed higher differential activity. Both the lncRNA cluster 3 network (Fig. 2.3) and lncRNA cluster 4 one (Fig. 2.4) contain network nodes identified as hVIPER features which retain correlation links from both lncRNA-hVIPER and lncRNA-PARADIGM comparisons.

2.2 Castration-Resistant Prostate Cancer

2.2.1 Introduction

Androgen Deprivation Therapy (ADT) is the standard of care for men with metastatic prostate cancer. However, ADT is not curative in this setting, and progression to lethal metastatic castration-resistant prostate cancer (CRPC) is nearly universal. The androgen receptor (AR) antagonist enzalutamide is one of the principal treatments for metastatic CRPC patients [11][122][123]. The majority of patients benefit from treatment with this agent [11][122][123]. However,

disease progression is inevitable, and very little is known about mechanisms that contribute to clinical enzalutamide resistance [44]. Seeking to clarify the molecular mechanisms that underlie enzalutamide resistance, we analyzed the genomes and transcriptomes of CRPC metastases obtained prior to treatment and at the time of disease progression in patients treated with enzalutamide.

2.2.2 Results

Men with metastatic CRPC were enrolled on the West Coast Dream Team (WCDT) biopsy protocol. We describe results from patients with enzalutamide-naïve CRPC who underwent both a baseline biopsy prior to enzalutamide treatment and a repeat biopsy at the time of progression and whose matched biopsies had sufficient tumor for molecular analyses. Importantly, all of the progression biopsies were performed prior to discontinuing enzalutamide.

A variety of organ sites were biopsied, but bone was the most common site for both pre-treatment and progression biopsies (Fig. 2.6). In six patients, the exact same lesion was biopsied at baseline and progression.

To determine the genomic landscape of these tumors, we performed exome sequencing on matched biopsy specimens from 16 patients for whom DNA was available using a targeted panel of 124 genes commonly altered in solid tumors. Beyond the AR, which is discussed below, the most common gene alterations at baseline were in P53 or PTEN, which matches a prior report [118]. In most matched pairs, there was significant concordance of mutations or copy number alterations, demonstrating that new alterations in the genes examined were not common contributors to acquired resistance.

Therefore, we sought to characterize AR gene alterations between the baseline

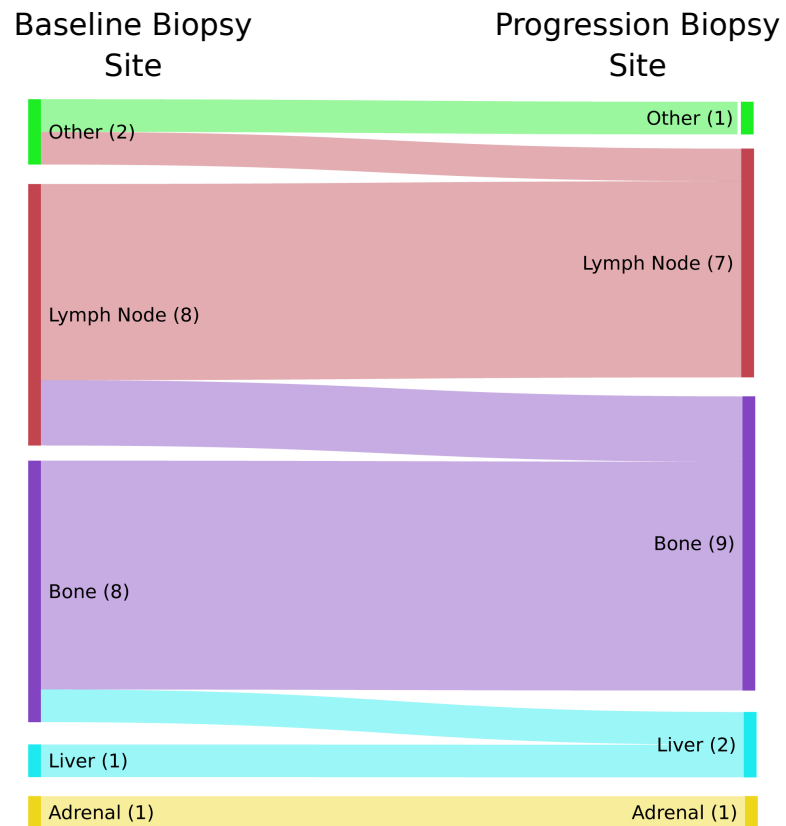


Figure 2.6: Site of biopsy for CRPC patients.

and progression samples. Increased AR copy number was common in baseline samples (9/16 patients), and AR remained amplified in all but one tumor at progression. We also examined the AR for mutations. Only one progression biopsy harbored an AR mutation—AR H875Y—that results in partial agonist activity by enzalutamide and confers broadened ligand specificity to non-androgen steroids [87]. Despite prior results demonstrating that AR is commonly altered in CRPC [118], there is no information about AR’s functional activity state in enzalutamide-resistant clinical samples. To quantify AR transcriptional function, we measured enrichment of a published AR gene signature [103] and a signature developed by the WCDT (see 2.2.3.1). For this analysis, we used 16 matched tumor pairs for which there was sufficient material for RNA-seq. There was a strong concordance between the results of these two signatures (Fig. 2.7, Spearman correlation=0.809, p-value=0.001).

Importantly, 11/16 patients were predicted to have reduced AR activity at progression based on the AR signatures. In two cases, loss of AR activity occurred in the setting of reduced AR mRNA expression, while nine others had reduced AR activity at progression without decreased AR expression. These results suggest that AR was less functional in the resistant vs. baseline samples.

Because AR function was predicted to be reduced in the majority of progression biopsies examined, we sought to identify other transcriptional regulators that contribute to enzalutamide-resistance. To achieve this, we used hVIPER on the 16 matched pairs with RNA-seq data. The top Master Regulators (MRs) predicted to have the greatest differential activity scores in the enzalutamide-resistant versus baseline samples are shown in Fig.2.8.

The MRs identified by hVIPER include several transcription factors and ki-

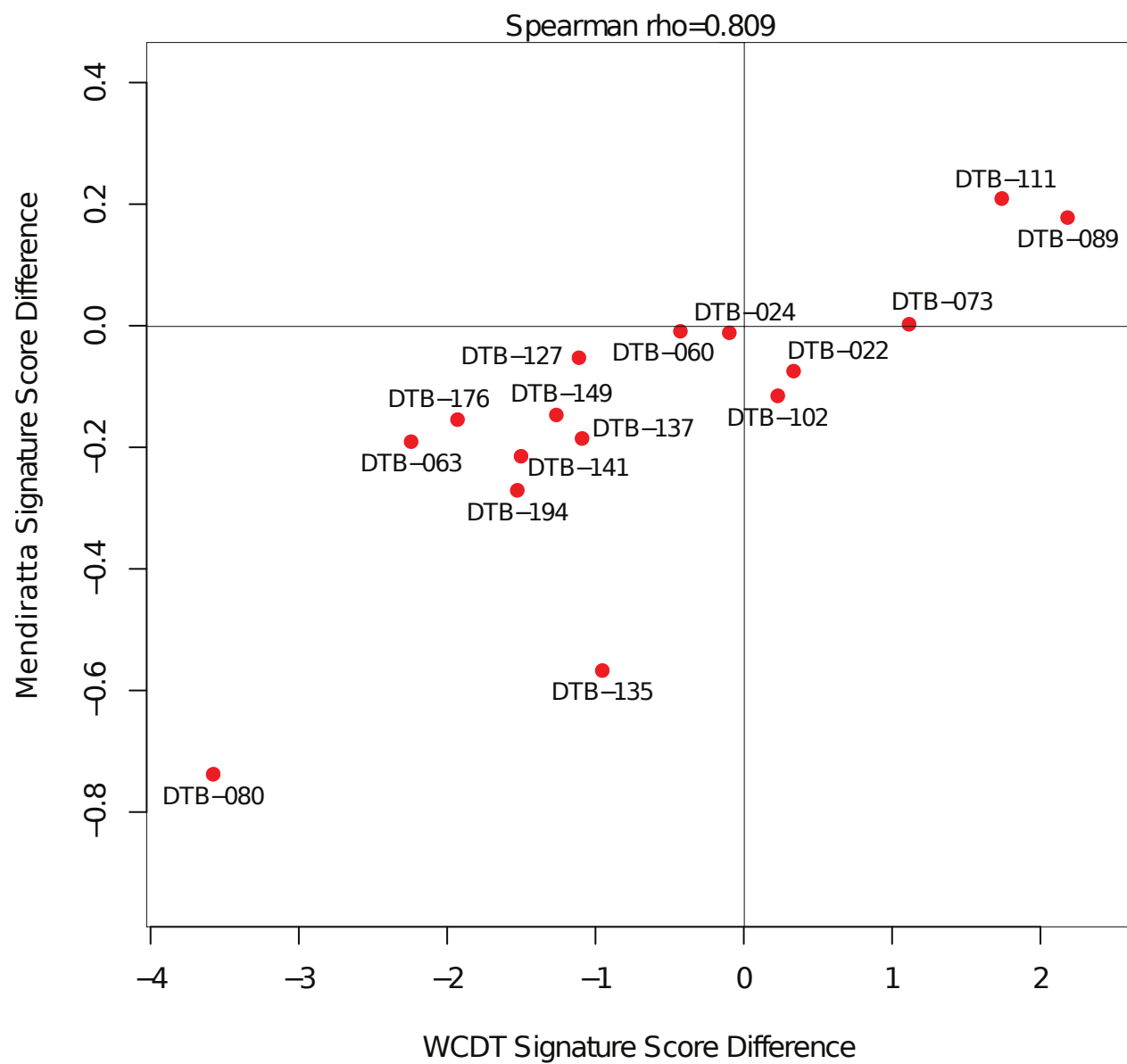


Figure 2.7: Concordance between the WCDT AR signature and the Mendiratta signature. Each point is a combination of differential (post- minus pre-treatment) scores for each of the two signatures.

Transcription Factors (FDR <= 0.2)						Kinases (FDR <= 0.2)		
Activated			Inactivated			Activated		
Regulon	NES	FDR	Regulon	NES	FDR	Regulon	NES	FDR
NR1I2	4.64	0.00063	DLX5	-5.12	0.000109	PRKCD	3.99	0.00542
NRF1	3.91	0.00541	RARG	-4.51	0.000761	CDK9	3.31	0.0263
FOXL2	3.78	0.00793	LEF1	-4.38	0.00104	PRKAA1	2.95	0.0548
POU3F2	3.73	0.00854	NR2F1	-4.17	0.00216	MAPK14	2.94	0.0548
CIC	3.51	0.0157	TBX2	-3.68	0.00931	CAMK2A	2.88	0.0554
MEIS2	3.34	0.0208	FOXA2	-3.47	0.0165	STK11	2.54	0.0949
BCL6	3.12	0.0362	KLF4	-3.38	0.0208	PRKDC	2.48	0.0949
GP1BB	3.11	0.0362	ESRRA	-3.35	0.0208	MAPK1	2.43	0.0949
ZZZ3	3.09	0.0362	ERF	-3.31	0.0221	MAPK8	2.41	0.0949
FOSB	3.02	0.0404	BCL3	-3.12	0.0362	GSK3B	2.35	0.105
TBP	2.96	0.0454	MYBL2	-3.08	0.0362	RPS6KB1	2.2	0.129
TCF4	2.9	0.0493	TBX5	-3.05	0.0385	CSNK1A1	2.06	0.162
PPARG	2.88	0.0497	DLX2	-3.01	0.0404	Inactivated		
CEBPB	2.84	0.0532	TBX3	-2.93	0.0463	Regulon	NES	FDR
PLEK	2.75	0.0644	ID3	-2.93	0.0463	AURKA	-3.47	0.0217
MTRNR2L1	2.7	0.0688	BARX2	-2.85	0.0532	CAMK2B	-2.75	0.0711
TAF1	2.7	0.0688	E2F2	-2.81	0.0563	PLK3	-2.48	0.0949
RORA	2.61	0.0797	LHX2	-2.75	0.0644	GSK3A	-2.46	0.0949
PRAME	2.58	0.0843	RUNX3	-2.7	0.0688	PRKAA2	-2.41	0.0949
CCDC116	2.54	0.0885	ID2	-2.65	0.0765	CDK2	-2.29	0.114
GATA4	2.51	0.0944	HOXA5	-2.61	0.0797	CHEK2	-2.2	0.129
MAFF	2.48	0.1	TWIST1	-2.61	0.0797	MAP3K5	-2.14	0.14
RELB	2.43	0.108	NCOA3	-2.58	0.0843	AURKB	-1.98	0.19
SP4	2.43	0.108	FOXC2	-2.54	0.0885			
FBXO31	2.39	0.118	TFAM	-2.43	0.108			
RAD21	2.33	0.134	SMC3	-2.39	0.118			
ELK4	2.27	0.148	MYCN	-2.37	0.123			
MAFK	2.26	0.152	MEIS1	-2.27	0.148			
CREB1	2.2	0.168	RUNX2	-2.24	0.155			
CTCF	2.13	0.192	DHRS2	-2.23	0.159			
GNB1L	2.13	0.192						
MAX	2.12	0.194						

Figure 2.8: Significant Master Regulators (FDR<0.2) discovered by hVIPER in paired-biopsy CRPC patients. NES is the differential activity score between pre- and post-treatment biopsies - positive scores (red) indicate higher differential activity in post-treatment biopsies, while negative scores (blue) correspond to relative inactivation.

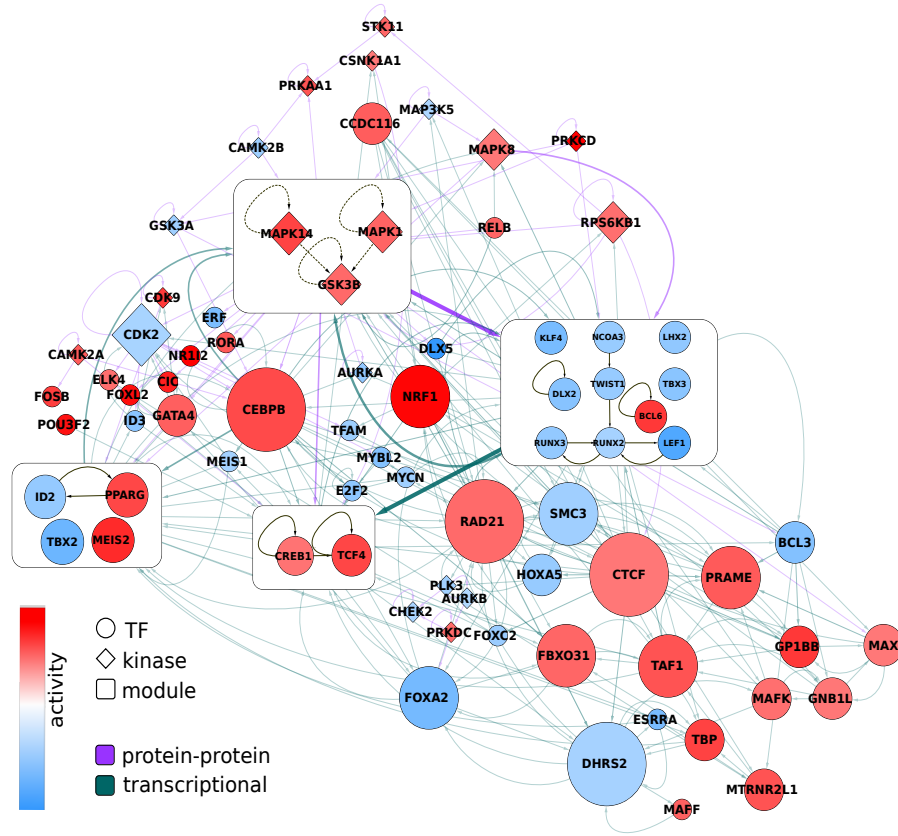


Figure 2.9: Full network of significantly differential Master Regulators. Nodes are colored by differential activity as in Fig.2.8. Shapes represent types of MR. Groups identified by MR Combinatorial Analysis are shown as multi-regulator nodes. Links are colored by type of interaction - green for transcriptional regulation; purple for post-translational regulation. Links to and from a module indicate regulatory interactions present between any of the individual module members and another MR/module. A module regulatory link is thicker if multiple module members interact with another MR/module - e.g. connecting a two-member module with a three-member could produce a maximum of six links among the Cartesian product of module members. If in fact only three of these are present in the MR regulome, the link thickness will be proportionately decreased. All links are plotted on a common thickness scale, with the thinnest lines corresponding to links between individual MRs.

nases that participate in a complex web of mutual interactions (Fig. 2.9). To summarize these results, I identified modules for which multiple regulators had related targets, suggested by VIPER Combinatorial Analysis (see 2.2.3.2). One such module included CREB1 and TCF4 as the 15 common targets had a differential activity (DA) score of 3.14 (Combinatorial Analysis $P < 0.002$; FDR=0.016) compared to the 138 individual targets of CREB1 (DA=2.2; $P < 0.03$; FDR=0.17) and the 417 individual targets of TCF4 (DA=2.9; $P < 0.004$; FDR=0.05). I then organized the interactions between modules and individual proteins to gain a high-level overview of the connections. Finally, I identified relevant cellular functions by aggregating Reactome pathways [46] with highest significant MR membership. This revealed a collection of circuits centered around the CREB1/TCF4 module involving several key processes such as immune-related responses, cell cycle, senescence/stress, nerve growth factor (NGF), and Wnt/Beta-catenin pathways (Fig.2.10). Several feedback circuits that operate through transcriptional and post-translational control logic were present. One such example is the Wnt-signaling feedback circuit that involves post-translational regulation by the P38 α /ERK complex on the CREB1/TCF4 and the BCL6/KLF4/LEF1 complexes.

Prior literature suggests an important role for CREB1 in castration-naïve or enzalutamide-sensitive prostate cancer, but CREB1's role in enzalutamide resistance is unknown [143]. Therefore, Joshi Alunkal performed CREB1 loss of function studies in enzalutamide-resistant CRPC cell lines. CREB1 RNAi suppressed viability of acquired enzalutamide-resistant models, suggesting functional importance of CREB1 (Fig.2.11A). Treatment with the CREB1 inhibitor 666-15 [90][172] recapitulated the effect of CREB1 RNAi (Fig.2.11B) and also reduced expression of luciferase reporters of CREB1 transcriptional function[90][172], further

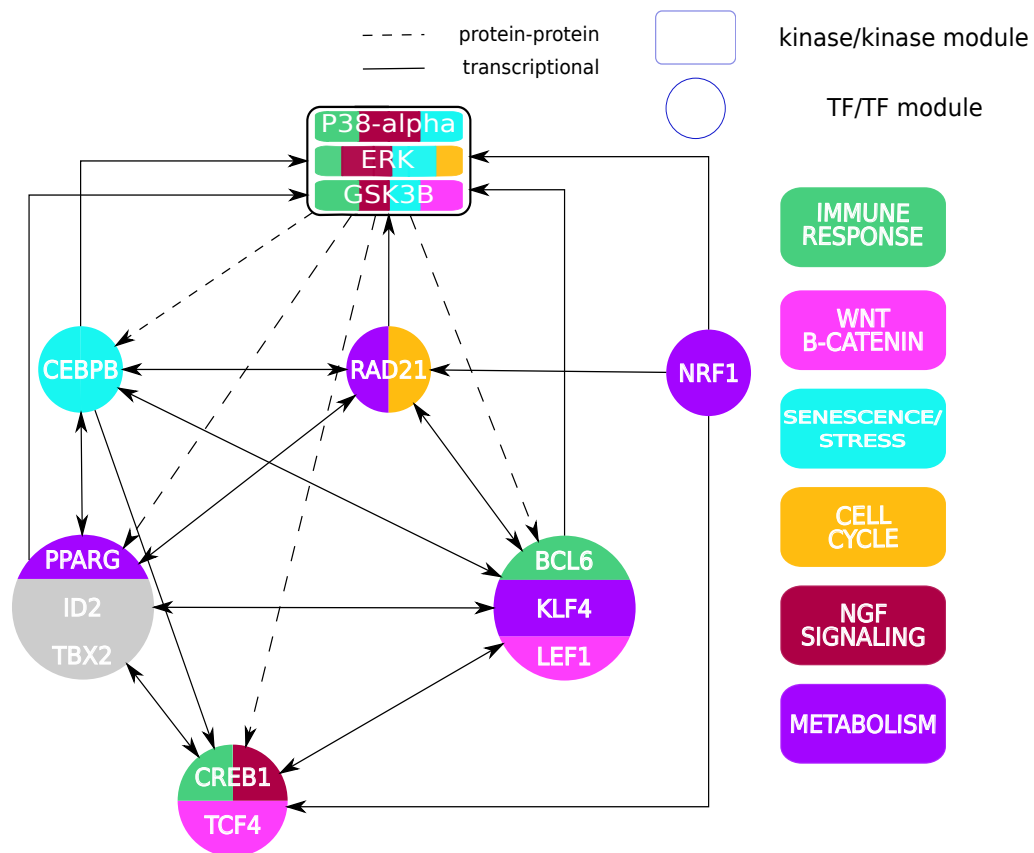


Figure 2.10: Main biological processes influenced by members of the four Regulator Combinatorial Modules characterizing pre- vs post- treatment CRPC (see section 2.2.3.2). Each regulator is involved in one or more processes as indicated by the regulator's coloring. Modules are shown as groups of multiple regulators.

demonstrating the relevance of CREB1 to enzalutamide resistance (Fig.2.11C).

Importantly, CREB1 is not the only drug-targetable MR implicated in our analysis. Relevant compounds that are FDA-approved for other conditions or in clinical trials include ketoconazole (blocks NR1I2), ralimetinib or ulixertinib (block MAPK1), and voruciclib (blocks CDK9) [162][165][41][109][141]. Exploring different combinations of CRPC-relevant MR targeting can increase our chances of mitigating enzalutamide resistance.

2.2.3 Methods

I used hVIPER (see section 2.1.4.2) with the regulator-target compendium described in 2.1.3.2 to estimate the differential activity of TF and kinase MRs in pre- and post- enzalutamide biopsies. I could not use generic t-test statistics for the gene signature required in multi-sample VIPER workflow (section 2.1.4.1) - had I done so, I would have been comparing MR activity differences across pre- and post-treatment cohorts instead of within individual samples. Instead, I created a ranked gene signature based on paired-sample t-test statistics. I also modified the permutation-based background model used for p-value/FDR calculations to prevent true sample biopsy pairs from occurring in permuted label sets (Fig.2.12).

Background models were computed by generating 1000 label permutations. Significance was evaluated by computing p-values against the background distribution and applying a BH FDR correction. Significant master regulators (FDR<0.2) are listed in Fig. 2.8.

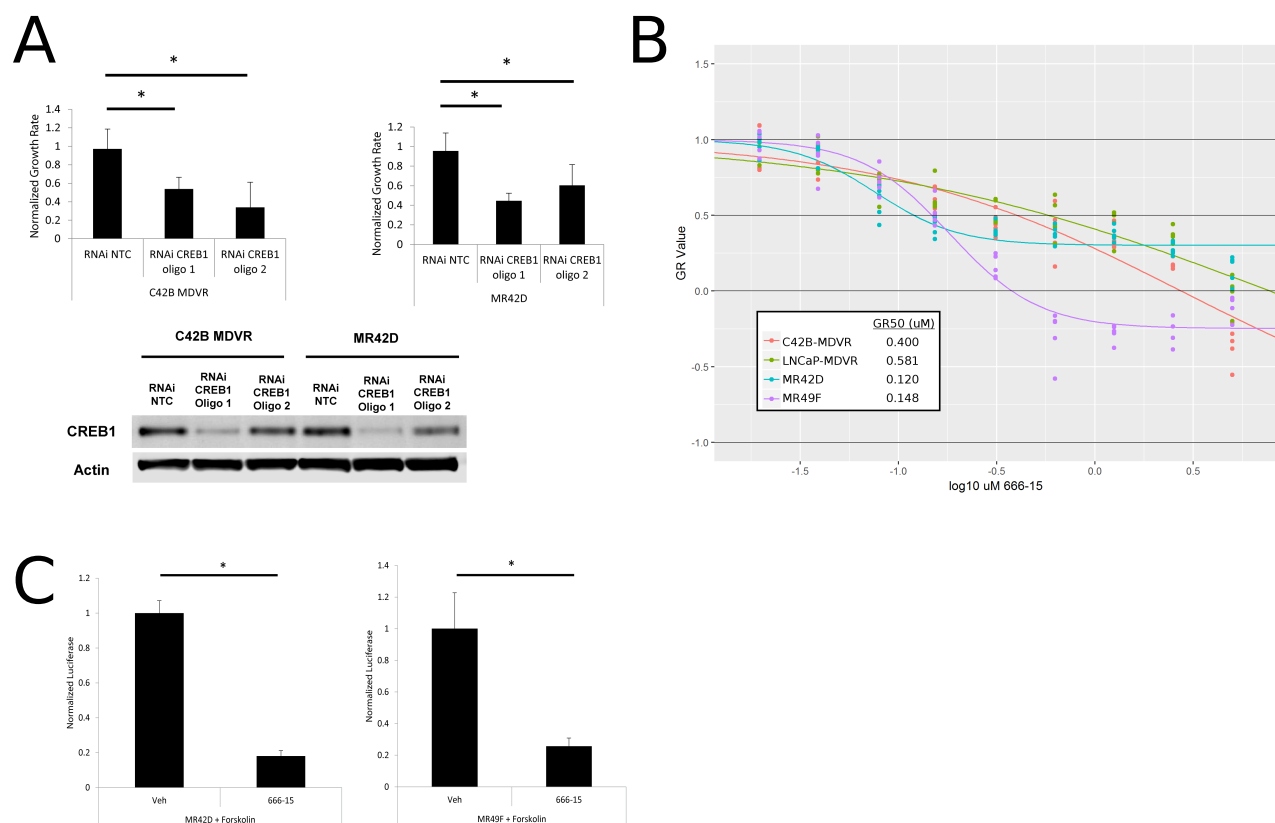


Figure 2.11: CREB1 RNAi and 666-15 experiments. A) CREB1 RNAi suppresses enzalutamide-resistant cell growth. The indicated cell lines were transfected with CREB1 RNAi or NTC RNAi. Cell viability was determined 120 hours later using fluorescent-labeled cell imaging. Protein lysates were harvested at 120 hours and probed with the indicated antibodies. B) CREB1 inhibitor 666-15 suppresses enzalutamide-resistant cell growth. The indicated cell lines were treated with dose escalation of the CREB1 inhibitor 666-15. Cell viability was quantified 72 hours later using fluorescently-labeled cell imaging. C) 666-15 treatment blocks the transcriptional function of CREB1. The indicated cell lines were transfected with a CRE-luciferase reporter plasmid. Cells were treated with 300nM of 666-15 or vehicle 24 hours post-transfection, and 10 μ M forskolin added 30 minutes post-treatment. Cells were harvested 6 hours post-treatment, and luciferase signal was quantified and normalized to total protein concentration. Experiments and data analysis performed by Joshi Alumkal.

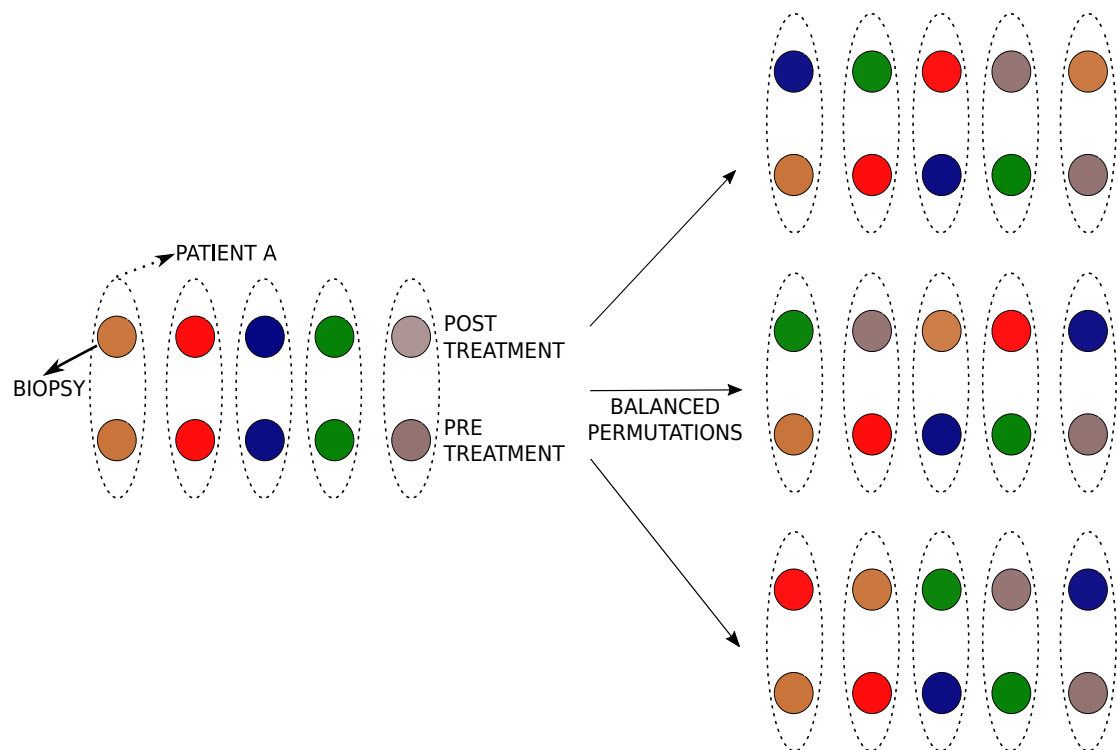


Figure 2.12: Balanced permutation procedure used in CRPC paired biopsy analysis. The true background model must exclude same-patient biopsies from matching up in the permuted folds.

2.2.3.1 AR signature evaluation

WCDT AR signature

I trained an AR signature on cell line profiles of pre- and post- AR ligand gene expression under various conditions [63][77][166][142][161][19]. Ligand status and expression data for 71 cell line profiles were downloaded from the Gene Expression Omnibus [10]. Three profiles were excluded due to low data quality (GSM288299, GSM288300, and GSM288301); the expression data for the rest was quantile-normalized by matching to the quantiles of an $exp(1)$ distribution. The expression data was further filtered to include only genes also measured in the WCDT CRPC cohort. I trained a generalized elastic-net classifier [135] (L_2 regularization coefficient = 1) to separate the AR-positive cell lines (ligand status "ON") from the AR-negative cell lines (ligand status "NO LIGAND"). To verify the performance of the classifier, I used leave-pair-out cross validation to measure the area under the curve (AUROC). The classifier had AUC of 1 (perfect accuracy for all cutoffs).

I then applied the trained model to WCDT expression data. To do so, I quantile-normalized the WCDT data as described above and took the dot product of normalized expression with the gene coefficients from the trained model.

Mendiratta AR signature

A published AR signature was taken from [103]. The signature provides weights for individual probes represented on the Affymetrix U133A platform. To convert the probe-level weights to the gene-level, I computed a median weight using all probes that map to a specific HGNC symbol. I then applied the signature to the WCDT samples. First, I found a common gene space by intersecting the set of HGNC symbols mapping to any of the signature probes with the symbols

mapping to the data used in this study. I then applied the signature to each WCDT sample by taking a dot product between the signature’s weights and a sample’s rank-transformed expression vector.

After scoring WCDT samples with both signatures, I computed the difference between signature scores in pre- and post- treatment biopsies (Fig. 2.7).

2.2.3.2 Circuit Generation CRPC

I constructed a connected and directed network that linked together significant master regulators (SMRs; $FDR < 0.2$). To do so, I collected all links in the TF and kinase regulons that involve at least one of the SMRs. The resulting “circuit” represents the connected subnetwork of individual MRs (Fig. 2.9).

I ran VIPER combinatorial analysis (VIPER CA)[\[6\]](#) to identify groups of MRs that are predicted to work together based on the behavior of their shared targets (VIPER CA is run separately for TFs and kinases). Specifically, VIPER CA retains any group of MRs for which the hVIPER activity score of the common targets regulon is higher than the activity score of the regulons’ union. I considered combinations of two and three MRs for all MRs that had a significant hVIPER result ($p\text{-value} < 0.05$) and found 44 combinatorial groups (CGs) - 36 TF and 8 kinase ones (Table 2.1). Because many of these groups are highly overlapping, (e.g. some were proper subsets of others), I defined a non-redundant set of merged Regulator Combinatorial Modules (RCMs) from these CGs. RCMs were created by merging any CGs having a minimum of 15 TF or 10 kinase common targets and sharing at least one MR. Four RCMs, each combining at least three TF or two kinase component CGs emerged from this analysis. Those include a BCL6-KLF4-LEF1 RCM consisting of 10 MRs and 17 CGs, a PPAR γ RCM consisting

of 5 MRs and 9 CGs, a CREB1-TCF4 RCM consisting of 5 MRs and 4 CGs, and a GSK3B-MAPK1-MAPK14 RCM consisting of three MRs and two CGs (Fig. 2.9). The RCMs were then filtered to only include MRs that appear in the SMR network described above. RCMs were connected to individual MRs or other RCMs if at least one regulator-target connection was present between any of the constituent MRs in a module.

To further distill the main biological processes characterized by the newly discovered SMR/RCM network circuit, I queried the Reactome [46] pathway database for pathways with the highest abundance of SMRs. I excluded generic functional pathways such as "developmental biology," "signal transduction," and "gene expression" as these provided little in the way of interpretation. Finally, I curated the highest ranked pathways into biological signaling abstractions based on the overlap in SMR membership among pathways. Figure 2.10 shows this stylized circuit representation for the four main RCMs we discovered.

Table 2.1: Combinatorial Analysis of significant MRs in CRPC.

Regulon	Size	NES	p.value	FDR	module type	included
MEIS2-TBX2	18	7.05	1.77E-12	6.91E-10	TF	1
BCL6-LEF1	25	6.16	7.18E-10	1.40E-07	TF	1
PPARG-ID2-TBX2	20	5.66	1.51E-08	1.96E-06	TF	1
KLF4-LEF1-RUNX2	19	5.55	2.85E-08	2.47E-06	TF	1
KLF4-LEF1	30	5.53	3.17E-08	2.47E-06	TF	1
KLF4-NCOA3	18	5.38	7.41E-08	4.82E-06	TF	1
RUNX2-TWIST1	40	4.96	7.17E-07	3.50E-05	TF	1
LEF1-RUNX2	26	4.81	1.53E-06	6.63E-05	TF	1
BARX2-ID2-MEIS2	18	4.42	9.74E-06	0.000317	TF	1
KLF4-RUNX3-RUNX2	17	4.18	2.91E-05	0.000742	TF	1
BARX2-PPARG-TBX2	18	4.17	3.04E-05	0.000742	TF	1
BCL6-KLF4	24	4.03	5.65E-05	0.0012	TF	1
ID2-MEIS2	18	4.02	5.91E-05	0.0012	TF	1
NCOA3-RUNX2	21	4.01	6.05E-05	0.0012	TF	1
PPARG-TBX2	23	4	6.33E-05	0.0012	TF	1
KLF4-RUNX3	29	3.99	6.49E-05	0.0012	TF	1
MAFF-MAFK	118	3.92	8.82E-05	0.00149	TF	0
DACH1-TAF1	16	3.91	9.19E-05	0.00149	TF	0
BARX2-PPARG-ID2	19	3.63	0.000285	0.00397	TF	1
KLF4-TWIST1	21	3.54	0.000397	0.00534	TF	1
FOXA2-TCF4	60	3.48	0.000493	0.0062	TF	1
DLX2-RUNX3-TWIST1	19	3.46	0.000542	0.0064	TF	1
BARX2-PPARG-MEIS2	18	3.37	0.000745	0.00831	TF	1
BARX2-PPARG	19	3.34	0.00084	0.00862	TF	1
TBX3-TWIST1	15	3.27	0.00107	0.0104	TF	1
CREB1-TCF4	15	3.14	0.00167	0.0156	TF	1
TAF1-TBP	104	3.14	0.00168	0.0156	TF	0
CCDC116-GP1BB	21	3.12	0.00182	0.0157	TF	0
RUNX3-TWIST1-LHX2	18	3.08	0.00207	0.0165	TF	1
LEF1-RUNX3-LHX2	18	3.01	0.00258	0.019	TF	1
DLX2-RUNX3	20	3.01	0.0026	0.019	TF	1
LHX2-RUNX3	18	2.94	0.00331	0.0229	TF	1
CCDC116-GNB1L	16	2.87	0.00406	0.026	TF	0
GATA4-GATA6	18	2.61	0.00915	0.0482	TF	0
CREB1-POU5F1	15	2.55	0.0109	0.0543	TF	1
HOXA9-POU5F1	15	2.07	0.0385	0.151	TF	1
MAPK1-PRKAA1	5	5.46	4.66E-08	4.24E-06	kinase	0
MAPK1-PRKCD	7	5.3	1.17E-07	5.34E-06	kinase	0
GSK3B-MAPK14	13	3.93	8.54E-05	0.00259	kinase	1
GSK3B-MAPK1	13	3.61	0.000303	0.00688	kinase	1
GSK3B-MAPK8	7	2.98	0.00286	0.0287	kinase	0
CDK2-GSK3B	6	2.95	0.00321	0.0287	kinase	0
CAMK2A-MAPK1	5	2.74	0.00606	0.0459	kinase	0
CDK2-MAPK1	8	2.42	0.0154	0.0728	kinase	0

Size = number of common targets to all MRs in a group. NES=inferred activity score. Included=whether MR group is retained (1) or not (0) for subsequent analysis, based on 1) are common targets sufficiently many (≥ 15 for TFs and ≥ 10 for kinases) and 2) are the member MRs part of the connected subnetwork defined by TF and kinase literature regulomes.

2.3 Conclusion

In this chapter I proposed a feature engineering approach (hVIPER) that extends previously existing methodology (VIPER). Both VIPER and hVIPER address a commonly encountered problem - most experiments generally provide genomic and/or transcriptomic data, but not protein activity measurements for key regulators. VIPER presents an elegant way to generate inferred TF protein activities from the expression readings of their targets. hVIPER goes a step further and provides inferred kinase activities, doing so in a prior knowledge-informed manner. I demonstrate how this extension can provide key insights into the biological processes defining poor-surviving UM subtypes as well as the development of enzalutamide resistance in CRPC patients. The latter case is a perfect example of how hVIPER can lead towards novel clinical therapies - the addition of kinase activities allowed for the discovery of an MR circuit of mixed transcriptional and post-translational regulation that would have remained obscured or only partially characterized if I used VIPER alone. Furthermore, kinases provide a much more accessible drug targeting avenue - even though our main finding was a TF (CREB1), we discovered several other targetable MRs that could potentially supplement or improve upon the efficacy of targeting CREB1. More importantly, our analysis provides a blueprint of how such targetable signaling circuitry can be discovered in other experimental settings, solely using expression data.

Chapter 3

Gene Essentiality Prediction via Multiple Pathway Learning

This chapter describes one of the co-winning methods of the DREAM9 gene essentiality prediction challenge [59]. I was the lead developer of Multiple Pathway Learning (MPL) and a co-first author of the DREAM9 marker paper [59]. The introduction, data description and analysis of combined results contain material attributable to all authors. Evan Paull, a Stuart Lab PhD alumnus, contributed the RF models used in the MPL-RF ensemble. The results specific to MPL are my own work.

3.1 Introduction

Although genetic alterations of human tumors have become increasingly well characterized, the application of this knowledge to clinical therapy has been limited. The promise of targeted cancer therapy requires both effective treatments and accurate, easily tested biomarkers to identify patient populations likely to re-

spond to those treatments. Genes that are required for the survival of tumor cells, but not for normal cells, may provide an opportunity for specific targeting. This category of essential genes creates preferential vulnerabilities only in the context of a tumor’s specific genetic and epigenetic background, but not in contexts lacking those alterations. For example, the PARP1 gene has been shown to be essential in tumors deficient in BRCA1 or BRCA2, thus making PARP inhibitors a candidate for treating selective subtypes of breast and ovarian cancers [18][48][43]. Therefore, a critical need exists to accurately predict differences in gene essentiality across a wide variety of cancer genetic subtypes from molecular features.

Large-scale functional screening of molecularly characterized cancer cell lines is a promising approach to generate pre-clinical hypotheses of tumor subtypes associated with sensitivity to functional perturbations. Several recent studies have identified small numbers of validated biomarkers of preferential gene essentiality based on statistical analysis of large-scale genetic screens [124][98][130][99]. However, no study has yet conducted a systematic analysis of modeling approaches designed to infer predictive models of relative gene essentiality of cancer cells from such functional screening projects. Such a study is not only a challenging scientific problem, but one that emphasizes the need for a community of scientists with a range of expertise in data generation, predictive modeling, and biologic interpretation.

The DREAM challenges (<http://www.dreamchallenges.org/>) aim to assess computational models, contributed by researchers across the world, to prediction tasks focused mainly on biomedical research problems. Participating research teams develop and fine-tune their predictive models during one or more model building rounds. In the final round, each research team submits their best models, and,

after the challenge is closed, an unbiased assessment using standardized metrics is performed on a blinded set of test data.

We used DREAM challenges to study the predictability of relative gene essentiality by leveraging Project Achilles, which has created one of the largest publicly available datasets containing genome-wide RNAi-mediated screens in molecularly characterized cancer cell lines [34]. We examined 149 cell lines tested for gene essentiality by a library of 98,000 shRNAs targeting 17,000 genes. To avoid teams predicting off-target effects of RNAi screens, relative gene essentiality scores were calculated using the DEMETER algorithm, which models out and corrects for off-target miRNA seed effects [154]. The challenge also utilized molecular feature data from these 149 cell lines (genome-wide gene expression and copy-number data, in addition to mutational profiling of 1,651 genes) from the Cancer Cell Line Encyclopedia [9]. Based on these data, participants were tasked with training a predictive model on a subset of 105 lines with both molecular feature and relative gene essentiality data and applying it to infer the relative gene essentiality values for the held out batch of 44 cell lines for which participants could only access molecular feature data (Fig.3.1). Predictions were scored by Spearman’s rank correlation coefficient for each gene between the measured and predicted relative gene essentiality scores.

3.2 Results

3.2.1 Overall Results - Subchallenge 1

The average ranks for all teams across 14738 prediction tasks are shown in Fig. 3.2 - the comparison metric for each task is the Spearman correlation between test labels and a method's predictions (average Spearman correlation for each team across all tasks is displayed in the top row). The scores of the top 4 teams displayed statistically improved performance from the fifth ranked team, but were not statistically significantly different from each other. A more stringent post-hoc test (nonparametric Friedman's test) determined that the top three teams achieved improved performance from the fourth ranked (red stars in Fig. 3.2, for more information see *Online Supplement* in [59]).

As a baseline, a lasso regression model was trained with the same training and test data and achieved mean Spearman correlation of 0.1528 (Fig. 3.3) - a result that was bested by 18/21 teams, suggesting that more sophisticated models can lead to a large boost in shRNA knockdown prediction accuracy. Furthermore, when the same baseline model was applied to 10,000 permutations of the test set (randomly shuffled cell line labels), the mean Spearman correlation of the permuted sets is centered around 0 with standard deviation of 0.0211 (Fig. 3.3), indicating that genomic data contains measurable signal that can at least partially correlate to the shRNA knockdown phenotypic effects.

3.2.2 Results - Multiple Pathway Learning

Our Multiple Pathway Learning - Random Forest ensemble (MPL-RF, described in 3.4.3) was the joint co-winner, and second highest performer overall, on

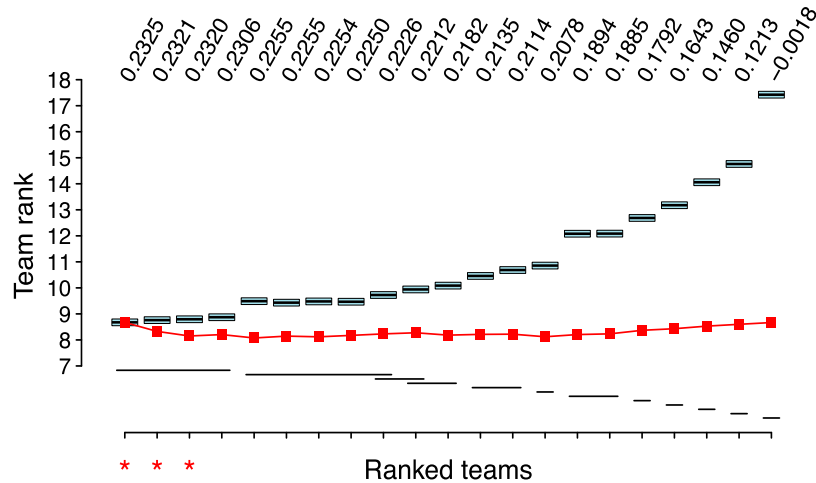


Figure 3.2: Team’s performances on DREAM gene essentiality prediction subchallenge 1. The average Spearman correlation across all predicted genes is shown at the top. Black lines in the middle of blue boxes (“team rank”) display the average rank of each team’s scores across 14,738 genes. Black lines below the boxplots summarize the Wilcoxon signed rank test results by dividing the teams into groups. Each group is denoted by a distinct black line, which contains the teams with no statistically significant difference. The red line displays the rank of an ensemble model constructed from all models with rank less than or equal to the corresponding point on the x axis. Red stars on the x-axis indicate the three joint co-winning teams, determined by non-parametric Friedman’s test - MPL-RF is the middle star. Figure taken from [59].

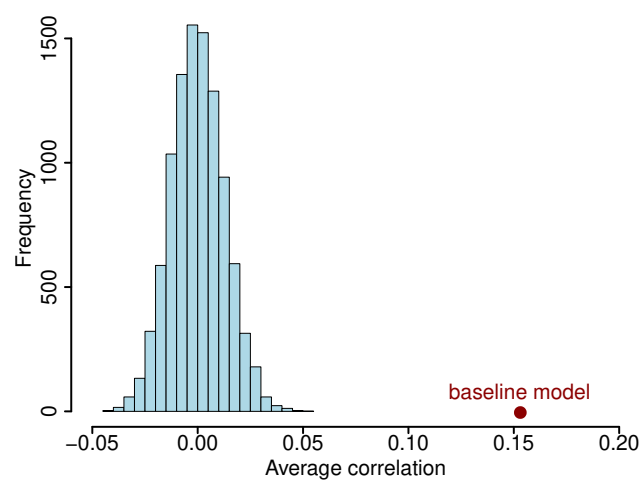


Figure 3.3: Mean Spearman correlation for a baseline lasso regression model (red dot) across all prediction tasks. The histogram represents mean Spearman correlations for a null distribution computed by randomly permuting the labels of the test set 10,000 times and applying the baseline model to the permuted test data. Figure taken from [59].

subchallenge 1 (0.2321 mean Spearman correlation, Fig. 3.2). Multiple Pathway Learning (MPL, described in 3.4.2) on its own was the fifth best performing model (0.2293 mean Spearman correlation, not shown in figure). As the average scores clearly indicate, predicting shRNA profiles is quite hard - most knockdown models produce surprisingly poor fit to test data. There are at least two reasons for such lack in performance. On one hand, gene expression was measured independently of the knockdown procedure and could therefore provide overall transcript levels only instead of the much more informative differential expression before and after RNAi. Since expression is generally the most informative data type, this leads to significant signal attenuation. On the other hand, shRNA knockdown is a notoriously noisy procedure with significant side effects - it is quite likely that the DEMETER score correction can only partially counteract that biological artefact. In fact, in their Dependency Map analysis [154] the authors focused on 769 knockdown profiles in which at least one depletion score was ≥ 6 standard deviations from the mean across all cell lines and all knockdown profiles.

Taking these 769 shRNA knockdowns as the high-confidence portion of the label set, we can re-evaluate the performance of the top methods on tasks with reliably measured labels. Because Dependency Map was based on ≈ 3 times more cell lines than what was made available in the DREAM challenge, there were only 426 gene profiles matching the 6σ criteria in the latter.

As expected, the high-confidence profiles produce significantly improved average Spearman correlations for all methods (Table 3.1) and those improvements are positively correlated with the number of 6σ measurements within each shRNA profile (Fig.3.4). Interestingly, when only high confidence profiles are considered, MPL (mean Spearman 0.3396) performs just as well as MPL-RF (mean Spearman

Table 3.1: Mean Spearman correlations of MPL, RF, and the 3 top performing methods from subchallenge 1 for high-confidence shRNA profiles. Individual correlation statistics for non-MPL DREAM9 methods downloaded from www.synapse.org (synapse ID syn2384331).

method	mean Spearman
MPL	0.3396 $\pm$ 0.0071
MPL-RF	0.3385 $\pm$ 0.0070
DREAM9-w2	0.3256 $\pm$ 0.0072
DREAM9-w1	0.3121 $\pm$ 0.0071
RF	0.3105 $\pm$ 0.0074

DREAM9-w1 is the top-ranked method overall in subchallenge 1, MPL-RF is the second best, and DREAM-w2 the third best.

0.3385) and significantly better than both of the other winning methods (p values of 1.7×10^{-13} and 0.0033 Wilcoxon signed-rank pairwise test, Fig. 3.5, Table 3.1).

3.2.2.1 Towards a Predictor-Knockdown Pathway Interaction Network

I further interrogated the MPL models to determine if there are any similarities between the pathways involved in predicting different shRNA knockdown profiles. I chose to focus on MPL and not MPL-RF models because of the considerations in the previous section as well as the significantly more difficult interpretation of the MPL-RF ensemble. Due to the nature of the non-linear Gaussian kernels representing each pathway (see 3.4.4.5), it is impossible to create a reverse mapping that would allow the assignment of importance weights to the original input features. Nevertheless, I can still extract pathway importance scores that quantify the contribution of each MPL pathway kernel to individual shRNA profiles. I refer to the pathways with non-zero weights in any MPL model as predictor pathways (PP).

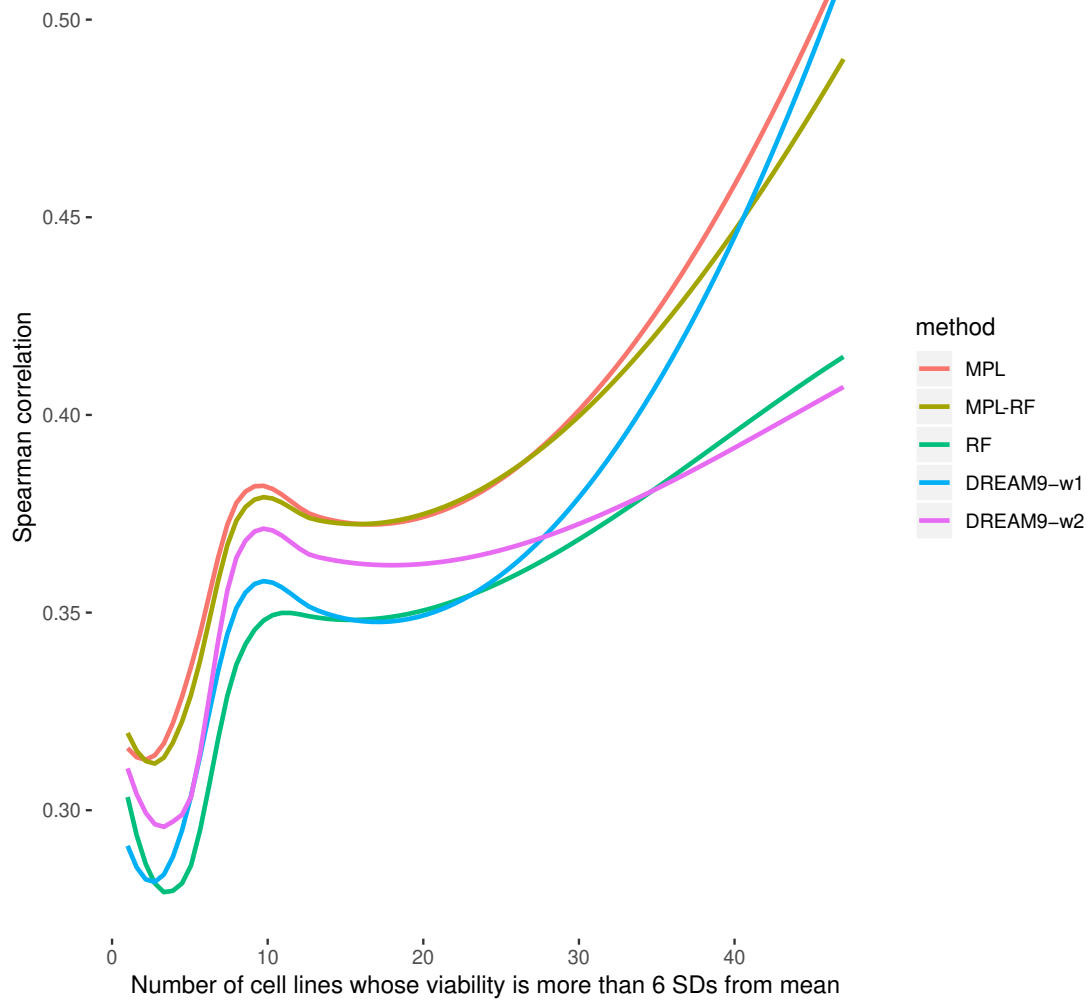


Figure 3.4: Relationship between number of 6σ measurements in an shRNA profile and the predictive accuracy of trained models for the best performers in subchallenge 1. The curves represent a smoothed approximation by local polynomial regression [27] computed via the R function *loess* with default parameters.

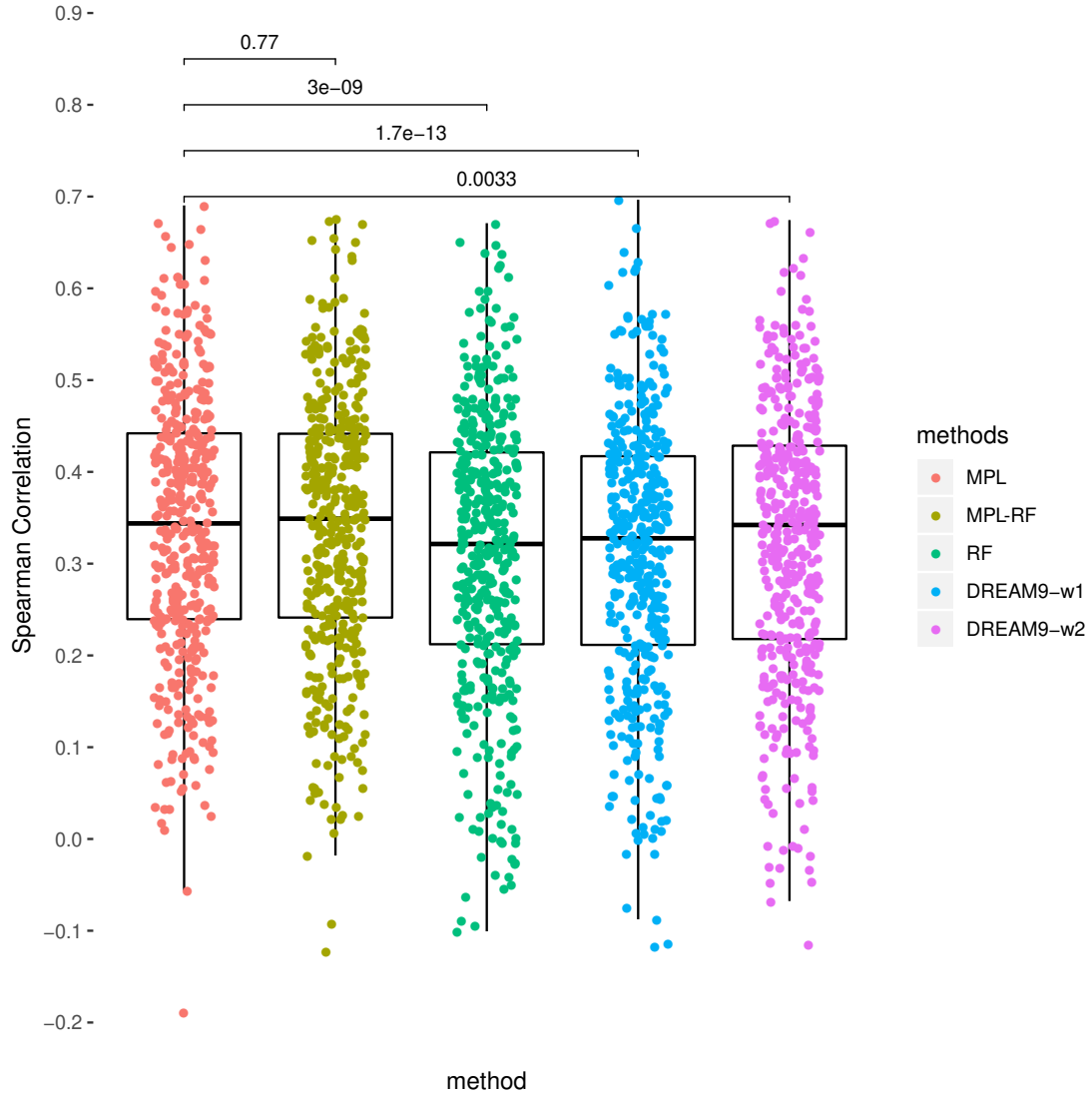


Figure 3.5: Spearman correlations for top subchallenge 1 methods on 426 high-confidence shRNA profiles. DREAM9-w1 is the top-ranked method overall in subchallenge 1, MPL-RF is the second best, and DREAM-w2 the third best. P-values for Wilcoxon signed-rank test pairwise comparisons.

On the other hand, even though some knockdown profiles are more robust than others, I would like to use as much of the shRNA data as possible. In particular, it is very important to characterize the biological processes (knockdown, or KD pathways) in which we can predict shRNA knockdown effects most accurately. Mapping PP to KD pathways can help define a pathway-pathway interaction network in which we can trace the effects of pathway dysregulation on related cellular processes. To that effect, I defined a gene's knockdown prediction score as the Spearman correlation between my test set MPL predictions and the actual knockdown measurements. I then applied Gene Set Enrichment Analysis (GSEA)[139] to determine which "canonical" pathways are enriched for genes with high knockdown prediction scores (see 3.4.4.1 for a definition of "canonical" pathways). A high enrichment score indicates that most of the genes in that gene set have high knockdown prediction scores and are thus well-characterized by the associated MPL model. The most enriched KD pathways are generally related to essential biological processes (e.g. ATP production, transcriptional control, cell cycle progression and proteasomal degradation, Table 3.2) whose disruption is often catastrophic for the cell. Many of the key regulators of cell behavior, such as TP53, RB1, VEGFA, CDK1/2, CHEK1, VHL and CCND1 appear in the most enriched portions ("the leading edge", see [139]) of each of the top 10 KD pathways (Fig. 3.6). This suggests that MPL models can provide a suitable avenue for analysis of disruptions in such key cellular processes.

We should point out that the leading edge genes for the top KD pathways are not a subset of the group of genes with robust shRNA knockdown profiles (see 3.2.2) - for example, TP53, RB1, and VEGF do not appear in the latter. Even so, MPL significantly outperforms both RF and the MPL-RF ensemble in predicting

the essentiality of leading edge genes - for the 101 such genes in the top 10 KD pathways (Fig. 3.6) MPL achieves mean Spearman correlation of 0.461 ± 0.007 versus 0.447 ± 0.008 for MPL+RF and 0.402 ± 0.010 for RF. Part of this improvement is due to selection bias - the top KD pathways are by definition enriched for the knockdown profiles MPL predicts best. Another reason, however, is that many of these genes are well studied, hence they have a higher proportion of relevant gene signatures in our gene set compendium. As a consequence, RF models are at a disadvantage because they do not incorporate such prior knowledge.

Table 3.2: Top 15 canonical pathways enriched for genes with highly predictive MPL models.

NAME	SIZE	NES	FDR
REACTOME FORMATION OF ATP BY CHEMIOSMOTIC COUPLING	12	1.92	0.14
REACTOME MRNA SPLICING MINOR PATHWAY	41	1.86	0.17
PID TOLL ENDOGENOUS PATHWAY	23	1.78	0.35
REACTOME G0 AND EARLY G1	22	1.77	0.30
BIOCARTA PROTEASOME PATHWAY	28	1.75	0.29
BIOCARTA LECTIN PATHWAY	11	1.74	0.29
BIOCARTA VEGF PATHWAY	27	1.72	0.32
BIOCARTA SKP2E2F PATHWAY	9	1.72	0.28
PID P38GAMMADELTA PATHWAY	9	1.70	0.30
BIOCARTA RB PATHWAY	12	1.70	0.27
REACTOME SCFSKP2 MEDIATED DEGRADATION OF P27 P21	50	1.69	0.27
BIOCARTA CELLCYCLE PATHWAY	21	1.68	0.29
REACTOME INTEGRATION OF PROVIRUS	8	1.67	0.30
REACTOME REMOVAL OF THE FLAP INTERMEDIATE FROM THE C STRAND	7	1.66	0.31
PID PLK1 PATHWAY	39	1.66	0.30

Results of GSEA enrichment analysis with MSigDB CP Canonical Pathways collection and the ranked vector of Spearman correlations between MPL predictions and shRNA knockdown measurements in 44 test cell lines. Size - size of pathway; NES - normalized enrichment score; FDR - False Discovery Rate q-value

Finally, we can harness the results of our GSEA KD pathway enrichment analysis to construct a PP-KD pathway interaction map. It is a bipartite graph

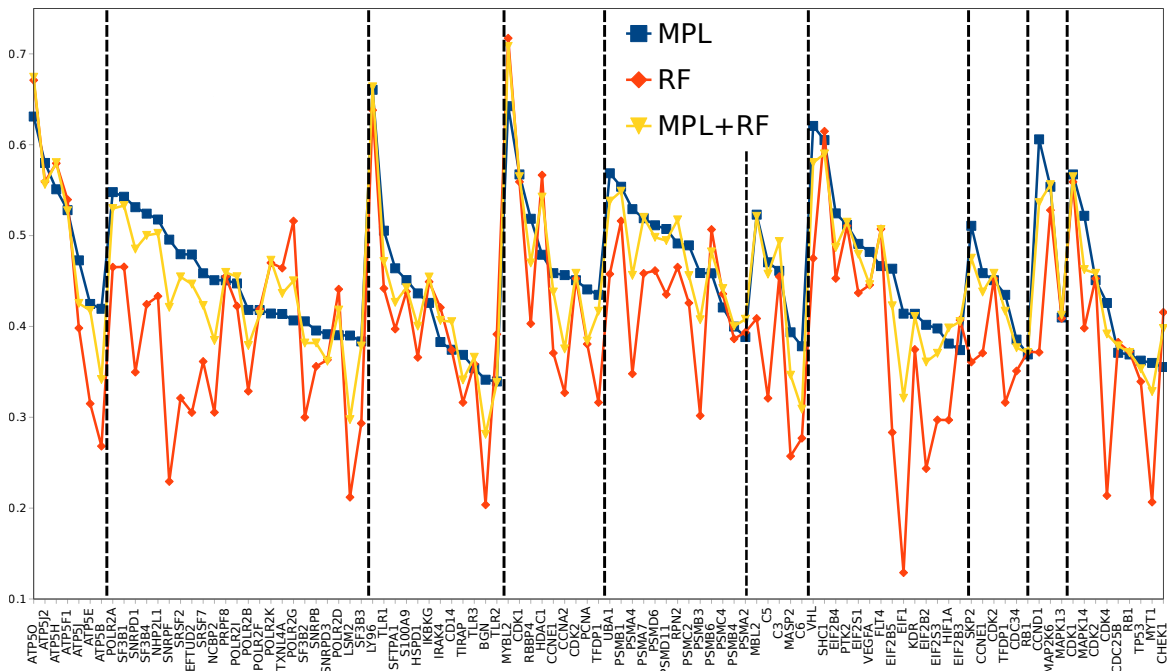


Figure 3.6: Leading edge genes for the top 10 canonical pathways from the GSEA analysis in Table 3.2. Spearman correlations between method predictions and each shRNA knockdown test profile shown for MPL, RF, and the MPL-RF ensemble. Vertical dotted lines mark boundaries between KD pathways. Pathways are ranked from left to right, with the leftmost the most enriched (REACTOME ATP FORMATION). Genes within pathways are ranked by the Spearman correlation of the MPL predictions to the true test set labels.

with KD pathways linked to PPs only and vice versa. Each edge has an associated weight that measures the importance of a PP to modeling the gene knockdowns in a KD pathway. We compute edge weights by averaging PP weights over a KD pathways' leading edge (Fig. 3.7). Fig. 3.8 shows the top 20 PPs for each of the top 20 KD pathways.

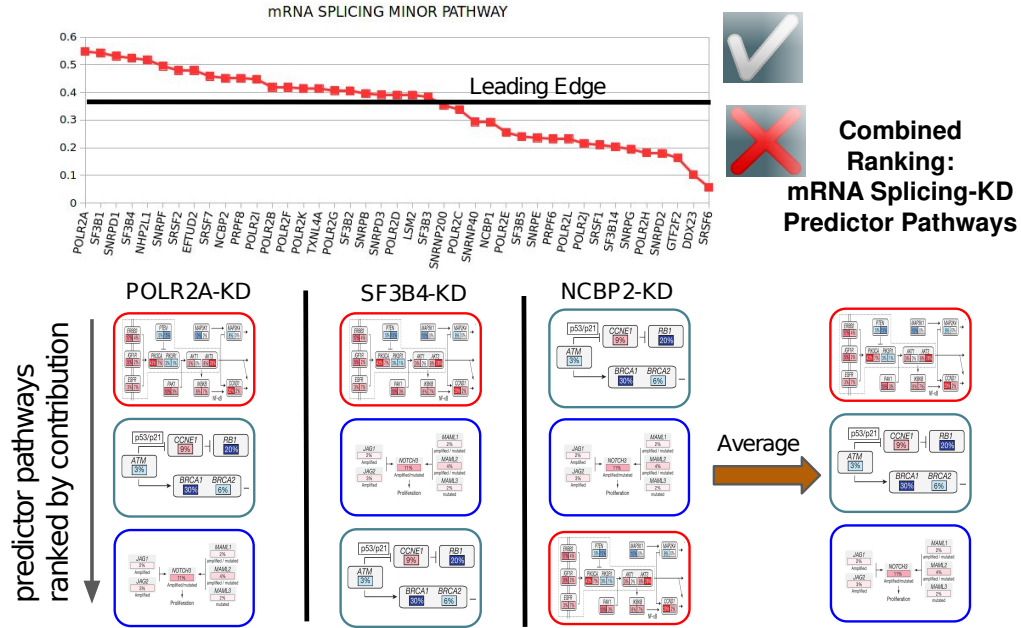


Figure 3.7: An overview of PP - KD pathway association strength score derivation, with mRNA Splicing KD pathway as an example. The MPL models for the leading edge gene are queried for their component non-zero weight PPs. If a PP has positive weights for both its expression and copy number kernels, they are summed to produce a composite weight. The PP composite scores are then averaged across leading edge models to give a measure of PP importance for modeling genes in a KD pathway.

As demonstrated by Fig. 3.8, many of the PPs remain very informative across all KD pathways (inner circle of PPs). On one hand, this is expected - the top KD pathways all represent essential cellular processes and disrupting their key components can lead to very similar cell viability profiles. Therefore, large degree of similarity in the PP weights for different models can and should occur. On the other hand, however, heavy overlap among top PPs for individual KD pathways make it very challenging to identify underlying biological processes that uniquely characterize the latter. We could focus on PPs that are shared among only a small subset of the KD pathways - for example, "NANOG Targets" is a top PP for the "G0 and early G1" and "P27" KD pathways while "UVC Late Response" is specific to the "P38 Gamma Delta" KD one (see outer PP circle in Fig. 3.8). Such an approach can be problematic for two reasons, however. Firstly, we don't know how much overlap these putatively KD pathway-specific PPs have with other more non-specific PPs. PP membership overlap can give us a crude estimate, but that can be quite misleading - individual features vary widely in predictive importance hence treating them equivalently can lead to biased conclusions. Secondly, just because a PP does not appear among the top PPs for a KD pathway does not mean it is completely non-informative for that KD. In fact, all of my models assign non-zero weights to most of the PPs in our compendium, with small differences in importance scores leading to large changes in PP ranking.

These problems are broadly applicable to the MPL-derived PP-KD network as a whole rather than just the subset of unique PPs. The highly concordant PP signatures for different KD pathways indicate that the variability in knockdown profiles is captured by subtle adjustments to the weights of the same PPs rather than by using distinct sets of PPs. In such circumstances, the availability of

feature importance scores is crucial to disambiguating the molecular processes predictive of each KD pathway. Similarly, the heavy degree of overlap among both PP and KD pathways makes it difficult to say with confidence that a PP is the most informative for a given KD pathway - it is possible that another biological process is more relevant, but its importance weight does not appear material because it is split among multiple highly overlapping PPs. Evaluating PP feature contributions can mitigate this issue as well. Therefore, a fully functional PP-KD map should include individual genomic features in addition to the PP and KD pathway entities.

3.3 Data

Challenge organizers provided a genome-wide shRNA knockdown screen in 149 cell lines from the Cancer Cell Line Encyclopedia [9]. Individual shRNA profiles for shRNAs targeting the same gene were combined into a consensus profile through DEMETER [154]. For more information on DEMETER settings and QC steps see [154] and the Supplementary Information in [59].

Molecular features (expression, copy number and mutation data) for the 149 cell lines were downloaded from <https://portals.broadinstitute.org/ccle>. All DEMETER profiles and genomic data are available from the challenge Synapse project (www.synapse.org, synapse ID syn2384331).

3.4 Methods

3.4.1 Random Forest

Random Forest is an ensemble learning algorithm whose component learners are decision trees trained on perturbed versions of the training data. Each decision tree is unique in two important ways - on one hand, its training set is a bootstrapped resample of the original data, with some samples appearing multiple times and others not at all (the latter are termed "out-of-bag", or OOB). On the other hand, at each node only a random subset of all attributes is considered for splitting. An optimal split produces two groups that minimize within-group variance (regression tasks) or Gini impurity (classification) [16]. Optimal splits are computed until all leaf nodes reach a pre-determined size (1 in the case of fully grown trees). Predictions for training data are computed by walking the OOB

samples along their respective trees and averaging the leaf OOB predictions for each sample across the forest:

$$\hat{f}(X_i) = \frac{1}{|OOB_i|} \sum_{m \in OOB_i} f^m(X_i) \text{ where} \quad (3.1)$$

$$OOB_i = \{t : X_i \in OOB(Tree_t); t = 1 \dots T\}$$

Predictions for new samples are computed in a manner identical to OOB training sample predictions, with the difference that averaging is done over the whole forest (new samples are OOB in all trees).

In our work we used the RandomForest R package [92] for all computations. We grew RF ensembles of 1000 CART trees [16], with the random subset of K features available for each node split set to $K = \sqrt{F}$ (as suggested in [92]), where F is the total number of features. Each tree was grown until leafs had no fewer than 5 samples (the default minimum node size for regression tasks in the *RandomForest R* package).

3.4.2 Multiple Pathway Learning (MPL)

MPL is a classification framework that integrates heterogeneous data sources and pathway information in a combined prediction model. Each gene set is kernelized (see 3.4.4) and the problem is reformulated as learning the linear combination of pathway-based kernels[136]. To avoid overfitting, we used Multiple Kernel Learning with elastic net regularization [145]:

$$\min_{\alpha \in \mathcal{R}^{NM}, b \in \mathcal{R}} \sum_{i=1}^N \mathcal{L}(y_i, \sum_{m=1}^M \sum_{j=1}^N k_m(x_i, x_j) \alpha_{m,j} + b) + \lambda_1 \sum_{i=1}^M \|\alpha_m\|_{K_m} + \lambda_2 \sum_{i=1}^M \|\alpha_m\|_{K_m}^2 \quad (3.2)$$

where N - number of samples, M - number of kernels, k_m is the kernel matrix for the m th kernel, $\alpha_m = (\alpha_{m,1} \dots \alpha_{m,N})^T$ are the sample-specific weights for the m th kernel, and $\|\alpha_m\|_{K_m} = \sqrt{\alpha_m K_m \alpha_m}$ is the respective kernel norm. Intuitively, the regularization terms control the trade-off between prediction accuracy and model parsimony. For more details, see 4.5.2 and 4.5.3 in the next chapter.

3.4.3 MPL-RF Ensemble

The MPL ensemble computes predictions by averaging the predictions from RF (Section 3.4.1) and MPL (Section 3.4.2).

3.4.4 Kernel construction

3.4.4.1 Pathway collections

I used MSigDB C2 (curated gene sets, 4722 sets) and C5 (GO gene sets, 1454 sets) [139] collections as sources for our gene set compendium. The C2 collection comprises both "canonical" pathways (CP, 1320 sets) as well as gene signatures for various chemical and genetic perturbations (CGP, 3402 sets). The C5 collection is derived from a flattened representation of the hierarchical network of terms in the Gene Ontology [5][29]. The C2 and C5 collections consist of 6176 gene sets in total. To minimize computing time, prior to kernelization I used several filtering steps to reduce the number of gene sets as well as the number of genes within

each gene set.

3.4.4.2 Variance-Based Expression Filter

We reduced the dimensionality of gene expression data by ranking genes based on variance and only retaining the top 25 % most variable ones. There were 4707 expression features left post-filtering.

3.4.4.3 Pathway Redundancy Filter

The gene sets in our compendium exhibit high degree of overlap. To mitigate that redundancy, I implemented the following size-based filtering:

1. The gene sets from the two MSigDB collections were ranked by size from largest to smallest.
2. The largest one was kept.
3. Every subsequent gene set was only included if it did not overlap more than 65% with any previously included gene sets (if the overlap exceeded 65%, the smaller gene set was discarded).

3.4.4.4 Copy Number TCGA PANCAN Filter

Copy number variation is often the result of large-scale events in which whole chromosome arms can be amplified or deleted. As a result, biological “driver” genes can be obscured by a (large) majority of “passenger” gene events. To mitigate this effect, we only used copy number data for 1752 genes enriched for functional genomic events in multiple tumor types of the TCGA PanCan12 cohort [26].

3.4.4.5 Features To Gene Set Kernels

Prior to filtering, I z-score normalized each feature in the training set.

For expression kernels, I first reduced gene set membership to genes that pass our variance-based filter and then I applied the pathway redundancy filter. I retained 2804 gene sets, ranging in size from 2 to 634 genes (median size 25, Fig. 3.9A), from which I constructed Gaussian expression kernels

$$K_m(x_i, x_j) = \exp\left(-\frac{\|e_i - e_j\|^2}{\sigma}\right) \quad (3.3)$$

where e_i, e_j are the expression vectors for samples i, j . Each expression vector consists of the expression measurements for genes in the m th filtered gene set.

For copy number kernels, I applied the variance-based filter and pathway redundancy filter as before, followed by the TCGA PanCan functional events filter. The intuition behind my approach is that I wanted to focus on copy number events that were not only a priori known to be involved in cancerogenesis but also have supporting expression changes in the cell line molecular data. I retained 2338 gene sets, ranging in size from 1 to 67 genes (median size 3, Fig. 3.9B), from which I constructed Gaussian copy number kernels

$$K_m(x_i, x_j) = \exp\left(-\frac{\|c_i - c_j\|^2}{\sigma}\right) \quad (3.4)$$

where c_i, c_j are the copy number vectors for samples i, j . Each CN vector consists of the CN values for the genes in the m th filtered gene set.

I applied the same σ to all kernels (EXP and CN). I chose sigma by exploring hyperparameter space around the mean length of all filtered expression gene sets (mean length=51.5, see also 3.4.5).

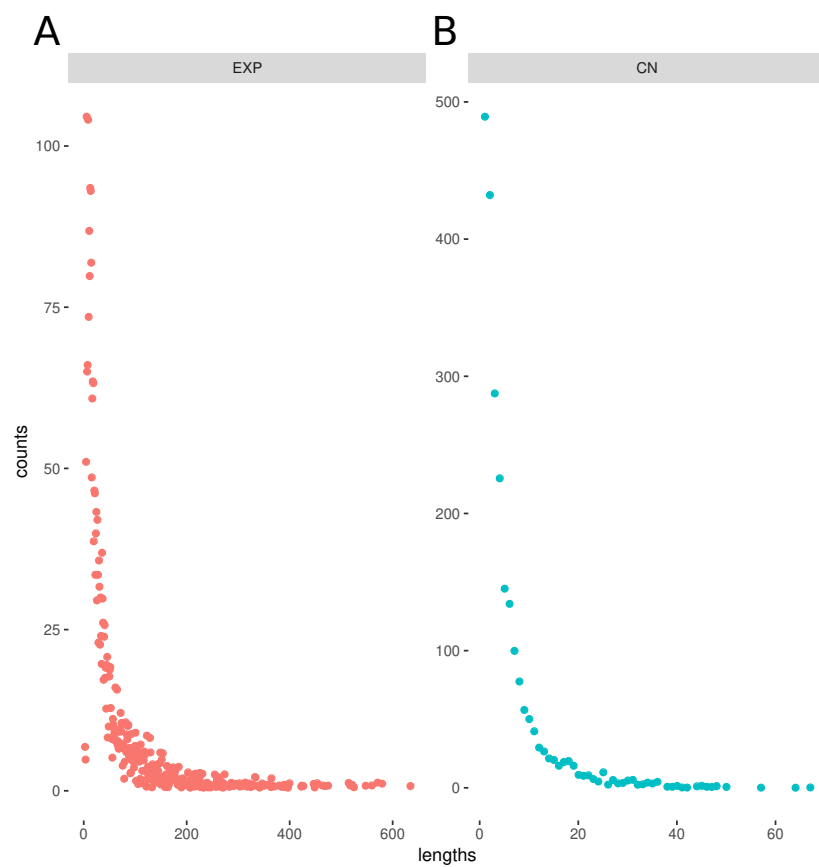


Figure 3.9: Distributions of A) Expression filtered and B) Copy number filtered gene set lengths. Filtering steps as described in text.

3.4.5 Hyperparameter selection

I tuned three MPL hyperparameters - the width of the Gaussian kernels σ and the parameters controlling elastic-net regularization (λ_1, λ_2) . I optimized them on the Phase 2 data of the DREAM9 challenge, with 66 cell lines available for training and 33 cell lines withheld as a test set [59]. To reduce computation time, I applied the same σ to all kernels and the same λ_1 and λ_2 to all prediction tasks. I considered values of 10, 50 and 100 for σ and λ vectors that induce sparse ($[\lambda_1 = 0.005, \lambda_2 = 0]$, few kernels with positive weights), dense ($[\lambda_1 = 0.0001, \lambda_2 = 1]$, almost all kernels have positive weights) and in-between ($[\lambda_1 = 0.001, \lambda_2 = 0.0001], [\lambda_1 = 0.0007, \lambda_2 = 0.0006]$) solutions. I also considered the utility of combining different data types by comparing models using only EXP kernels with ones trained on both EXP and CN kernels (both sets derived with the filtering steps described in 3.4.4). Finally, I evaluated the relative contributions of the three main groups of gene sets (CP, CGP and GO) in our gene set compendium.

I used a univariate hyperparameter selection approach - first I picked the best λ 's, followed by the best σ , the best data type combination, and the best gene set compendium, with each successive tuning run using the optimal hyperparameters from preceding runs.

My tuning experiments suggest that models in which most kernels receive non-zero weights tend to outperform ones with sparser weight vectors (Table 3.3). Similar conclusions have been previously reached from the analysis of non-bioinformatics data - sparser solutions often generalize less well and are frequently outperformed by baseline models such as the sum of uniformly weighted kernels [81].

The kernel width σ shows less effect on model accuracy, with values closer to

Table 3.3: Mean Spearman correlation for different (λ_1, λ_2) values

(λ_1, λ_2)	mean Spearman
(0.005,0)	0.1711 ± 0.0016
(0.001,0.0001)	0.1704 ± 0.0016
(0.0007,0.0006)	0.1748 ± 0.0016
(0.0001,1)	0.1849 ± 0.00156

Standard deviation is computed by dividing the sample sd by the square root of the number of observations (14738 tasks).

the mean length of the kernelized gene sets providing slightly better performance (Table 3.4). The inclusion of CN kernels in addition to EXP ones boosted performance to a larger extent than finding the optimal kernel width, but less so than optimizing MPL regularization parameters (Table 3.5).

Finally, the addition of information-rich gene sets provided the biggest accuracy gains (Table 3.6). Interestingly, the most highly curated collection (canonical pathways, CP) did not produce the most accurate models - both GO and CGP contained valuable orthogonal information that improved model fit. This is important for two reasons - on one hand, it highlights the fact that the best characterized regions of the human gene interaction network (CP) cover only a small portion of it, which is why exclusively focusing on them can lead to biased interpretations. On the other hand, curated signatures from prior experiments (CGP) can help shed light on the regions of the interaction network that CP does not cover. This is encouraging because experimental gene signatures are much easier to obtain - if more of them are made available to MPL, either by better aggregation of currently existing data or by incorporation of future experiments, model accuracies will likely improve.

Table 3.4: Mean Spearman correlation for different σ values

σ	mean Spearman
100	0.1849 ± 0.00157
50	0.1870 ± 0.00157
10	0.1866 ± 0.00159

Standard deviation is computed by dividing the sample sd by the square root of the number of observations (14738 tasks).

Table 3.5: Mean Spearman correlation for different data type combinations. Both EXP and CN were filtered as described in the text.

Data Type	mean Spearman
EXP	0.1870 ± 0.00157
EXP+CN	0.1926 ± 0.00156

Standard deviation is computed by dividing the sample sd by the square root of the number of observations (14738 tasks).

Table 3.6: Mean Spearman correlation for different gene set collections

Collection	mean Spearman
CP	0.2098 ± 0.00131
CP+GO	0.2192 ± 0.00131
CP+GO+CGP	0.2291 ± 0.00132

Standard deviation is computed by dividing the sample sd by the square root of the number of observations (14738 tasks).

3.5 Conclusion

This chapter introduces MPL - a kernel-based framework for the integration of pathway information and heterogeneous data types. MPL models achieve superior predictive accuracy in challenging regression tasks such as predicting cell line shRNA knockdown viability from molecular features. They provide a straightforward interpretation of the relative importance of a gene set to a particular prediction task. MPL can therefore be viewed as a way to blend together model learning and gene set enrichment analysis.

As all frameworks, however, MPL comes with some trade-offs. On one hand, it is not equipped to capture cross-data type feature-feature interactions as each data modality is kernelized separately. On the other hand, it cannot assess the relevance of individual features to modeling performance. The latter is particularly evident in my attempts to construct a pathway-pathway dependency map. Without a measure of feature importance it is extremely difficult to disambiguate the relevance of highly redundant gene sets - different pathways might share the same key driver features, but they will appear as distinct entities in the dependency map. This can confound the interpretation of PP-KD pathway pairs as the same biological process can be represented by multiple predictor gene sets. Both of these challenges are addressed by AKLIMATE - a novel algorithm I introduce in Chapter 4.

Chapter 4

AKLIMATE: Algorithm for Kernel Learning with Approximating Tree Ensembles

As we saw in Chapter 3, Multiple Pathway Learning (MPL) provides a very powerful framework for the integration of different genomic data types with pathway information. However it suffers from two distinct shortcomings. On one hand, it has no sound way of assigning predictive importance to the features within each pathway. On the other hand, the data integration model of MPL does not allow for feature-level cross-data type interactions within a pathway. To address these challenges, we introduce AKLIMATE.

4.1 Introduction

High-throughput data have become ubiquitous in the study of biological phenomena. Genome-wide structural variant estimates are available for thousands

of samples [140]. The exome is even better characterized, particularly in cancer studies, where tens of thousands of patients have copy number, mutation, methylation, mRNA, miRNA or protein levels measurements [66]. In addition, the drop in sequencing cost has led to a proliferation of experiments that capture multi-omic profiles under a plethora of conditions and perturbations. They produce feature sets that characterize differences between experimental conditions or provide wet lab validation for interactions at various steps of the Central Dogma. Such "prior knowledge" accumulates in an ever-growing number of databases [21][93][37][177][113]. However, the integration of prior knowledge with heterogeneous genomic measurements in an interpretable model remains a formidable challenge that is still not fully solved.

Scores of methods have been developed to address some aspects of the three main challenges (prior knowledge interrogation, heterogeneous data interrogation, and ease of interpretation), both for supervised and unsupervised analysis (for recent reviews, see [116][95][69]). Several common approaches emerge for supervised learning of sample phenotypes. A popular way of increasing interpretability is to train models with regularization terms that constrain the number of included features - sparse models are considered easier to interpret than ones containing thousands of variables. Common regularization penalties are the lasso [149] and the elastic net [176]. More sophisticated regularization schemes can control the model behavior at the feature set level - e.g. the group lasso (GL) [175] and the overlap group lasso (OGL) [72] - thus allowing the incorporation of prior knowledge in the form of feature set definitions. However, both have drawbacks that make them less suitable for biological pathway analysis where genes often participate in multiple processes. GL requires that if a feature has a coefficient of

zero in one group, it must necessarily have zero coefficients in all other groups. OGL tends to assign positive coefficients to (unions of) whole groups - this is problematic when the pathway definitions used to construct the feature sets are noisy or not appropriate for the learning task (e.g. using non-tissue-specific lists of transcription factor targets).

Network-regularized methods [135][91] use gene-gene interaction networks as regularization terms on the L_2 -norm of feature weights. While they do use pathway-level information and are straightforward to interpret, they focus on an individual data type. Multiple Kernel Learning (MKL) [7][115][145][57][129] approaches can incorporate heterogeneous data by kernelizing each set of features and learning a linear combination of the kernel representations. To prevent overfitting MKL methods generally include a regularization term - e.g., L_1 -norm [115] or the elastic net [145]. Prior knowledge can also be integrated by constructing individual kernels from a pathway's member features within each data type [59][57][129]. Indeed, MKL methods with prior knowledge integration [59][32] have won several Dialogue on Reverse-Engineering Assessment and Methods (DREAM)[137] challenges. Nevertheless, MKL suffers significant drawbacks when the contributions of input features need to be evaluated - except in trivial cases it is generally impossible to assign importance to the original features once the method is trained in the kernel function feature space. In addition, feature heterogeneity necessitates the construction of separate kernels for each data type, limiting the ability of MKL to capture cross-data-type interactions.

All these methods can in turn be used as components in more complex ensemble learning models. Ensembles combine predictions from multiple methods into a more robust, and often more accurate, "wisdom of crowds" final prediction.

The simplest and most common ensemble technique is averaging the predictions of component models. Averaging over uncorrelated models or models with complementary information can improve performance - it is one of the main reasons for the emergence of collaborative competitions such as the DREAM challenges [97][32]. In fact, such ensemble methods often win DREAM challenges [23] or outperform competitors in genomic prediction tasks [79]. While they provide a boost in predictive accuracy, their interpretation is quite challenging - due to the need for minimally correlated component model predictions, ensembles often combine different model types, making the computation of input feature importance impossible in the large majority of cases. A notable exception is the Random Forest (RF)[15] - an ensemble of decision trees that has been widely adapted in bioinformatics analysis.

A more general approach to combining the predictions of multiple models is model stacking [169]. In stacking base learner predictions are given as input to a stacked model which produces the final predictions. To reduce overfitting and improve generalization, base learner predictions are generated via a cross-validation scheme in which a sample prediction is calculated by a model trained on all folds but the one including the sample in question. Importantly, if the stacked model is a weighted combination of the base models (a Super Learner), it is asymptotically guaranteed to perform at least as well as the best base learner or any linear combinations of the base learners (including average and convex combinations) [159][111]. Stacked models exhibit identical pros and cons as standard ensemble models, with diminished interpretability traded off for increased accuracy. In some cases, however, interpretability can be tractable - [164] uses base RF learners with a stacked least square regression to compute a final weighted average of

the predictions from the base RFs. Although not examined in the original article, we demonstrate that a similar setup can lead to an intuitive derivation of feature importance scores. Furthermore, their approach scores well in terms of multi-omic integration. However, it incorporates no prior knowledge and, similar to the drawbacks of pathway-based MKL, the segregation of data types in the base learner stage excludes cross-data-type interactions.

In this study, we introduce the Algorithm for Kernel Learning with Integrative Modules of Approximating Tree Ensembles (AKLIMATE) - a novel approach that combines heterogeneous data with prior knowledge in a model that is easy to interrogate on the feature as well as feature set level. AKLIMATE harnesses the advantages of RFs (native handling of continuous, categorical and count data, invariance to monotonic feature transformations, ease of feature importance computation), MKL (intuitive integration of overlapping feature sets), and stacked learning (improved accuracy) while avoiding many of their shortcomings. It relies on two major computational insights. First, summary statistics of decision trees within an RF model can be used to compute a kernel similarity matrix (RF kernel) that is data driven yet capable of capturing complex non-linear feature relationships. Second, if an RF model is trained with only the features corresponding to a biologically relevant feature set, converting that model to an RF kernel allows us to handle many different kinds of heavily overlapped prior knowledge groups without the undesirable side effects of GL/OGL. What is more, the RF kernel naturally re-weights the input features so that more informative ones make bigger contributions to kernel construction - this property is key when dealing with noisy or partially relevant feature set definitions.

4.2 Results

We evaluated AKLIMATE performance on various prediction tasks - microsatellite instability in endometrial and colon cancer, survival in breast cancer, and shRNA knockdown viability in cancer cell lines. We benchmarked AKLIMATE against comparable methods that have performed well in recent DREAM challenges. We chose both classification and regression tasks as well as various levels of data availability - a single data type, multiple data types (including inferred data), or multiple data types with clinical information.

4.2.1 Microsatellite Instability

We first tested AKLIMATE on predicting microsatellite instability in the colorectal (COADREAD) and endometrial (UCEC) TCGA cohorts. Microsatellite instability (MSI) arises as a result of defects in the mismatch repair machinery of the cell. Tumors with MSI (often accompanied by higher mutation rate) represent a clinically relevant disease subtype that is associated with better prognosis. MSI is also an immunotherapy indicator as such tumors produce more neoantigens.

We used expression data and MSI annotations for the COADREAD and UCEC TCGA cohorts. The UCEC cohort consists of 326 patients, of which 105 exhibit high microsatellite instability (MSI-H) and the remaining 221 are classified as either low (MSI-L) or stable (MSS). The COADREAD cohort includes 261 samples, with 37 MSI-H and the remaining 224 classified as either MSI-L or MSS. In both tumor types, methods were trained to distinguish MSI-H patients from MSI-L+MSS patients. 50 stratified partitions of 75% training and 25% test were created for each cohort and methods were trained using only samples from the training split. We then computed area under the ROC curve (AUROC) for each set

of test set predictions. We chose this problem because MSI can be predicted with high accuracy from expression alone [57], giving us a straightforward benchmark for AKLIMATE performance with a single feature type in a binary classification setting.

We compared AKLIMATE to Bayesian Multiple Kernel Learning (BMKL)-BMKL performed well in several DREAM challenges, in particular winning the NCI-DREAM Drug Sensitivity Prediction Challenge[32]. Furthermore, its pathway-informed extension [57] shares several similarities with AKLIMATE - it is a multiple kernel learning method that operates on pathway-derived kernels. In particular, BMKL uses expression-based Gaussian kernels computed on features from the PID pathway collection [121]. We tested four versions of BMKL - sparse single-task BMKL(SBMKL), dense single-task BMKL(DBMKL), sparse multi-task BMKL (SBMTMKL), and dense multi-task BMKL (DBMTMKL). Sparse BMKL models are the focus of [57] - they use sparsity-inducing priors to train models with few non-zero kernel weights (we used the hyperparameters specified in [57]). We added DBMKL and DBMTMKL to the comparison because dense MKL models (almost all kernels receive non-zero weights) tend to produce higher predictive accuracy in many experimental settings (e.g. [150]). Their parameters are identical to the ones for the sparse models except for $(\zeta_\kappa, \eta_\kappa)$, which are set to $(999, 1)$ in the dense models ($(\zeta_\kappa, \eta_\kappa) = (1, 999)$ in the sparse ones). Finally, the single-task models train separately on the UCEC and COADREAD cohorts, while the multitask versions learn MSI status on the two TCGA cohorts jointly. All train/test splits are matched across methods, including multi-task ones. All BMKL models use 196 PID gene sets; all model parameters, kernel computations and data filtering steps match the setup described in [57]. To control for

the pathway sets used when comparing the methods, we created a reduced version (AKLIMATE-reduced) that was restricted to the same set of input PID sets available to BMKL. The unrestricted, full model of AKLIMATE is referred to as AKLIMATE in this comparison.

AKLIMATE predicted MSI status in the UCEC cohort significantly better than AKLIMATE-reduced or any of the BMKL models (Fig.4.1a). In particular, AKLIMATE achieved a mean AUROC of 0.962 compared to 0.938 for AKLIMATE-reduced ($P < 4.1e-08$; Wilcoxon signed-rank test), suggesting a predictive benefit to using AKLIMATE's larger collection of gene sets. A larger gene set collection is both more likely to contain sets derived specifically to describe the MSI process and more flexible in terms of the possible combinations of component gene sets. Indeed, the most informative feature set according to AKLIMATE was "MSI Colon Cancer" (16.9% relative contribution to model explanatory power) - a gene expression signature for MSI-H vs MSI-L+MSS in COADREAD cohorts [74] (Fig.4.1b). Other highly ranked gene sets were also related to MSI. For example, the next two most informative sets were "GO DNA Binding" (4.55% relative contribution) and "REACTOME Meiotic Recombination" (2.4% relative contribution), both of which are strongly relevant to DNA mismatch repair (MMR). Of particular note, MLH1 was the top-ranked single feature in AKLIMATE (26.7% relative contribution) and was present in all of the top ten ranked pathway kernels (Fig.4.1b). MLH1 is a key MMR gene involved in meiotic cross-over [70]. Loss of MLH1 expression, usually through DNA methylation, is known to cause microsatellite instability.

These results demonstrate that AKLIMATE is able to pinpoint individual causal genes as it sifts through thousands of pathways. In contrast, the meta-

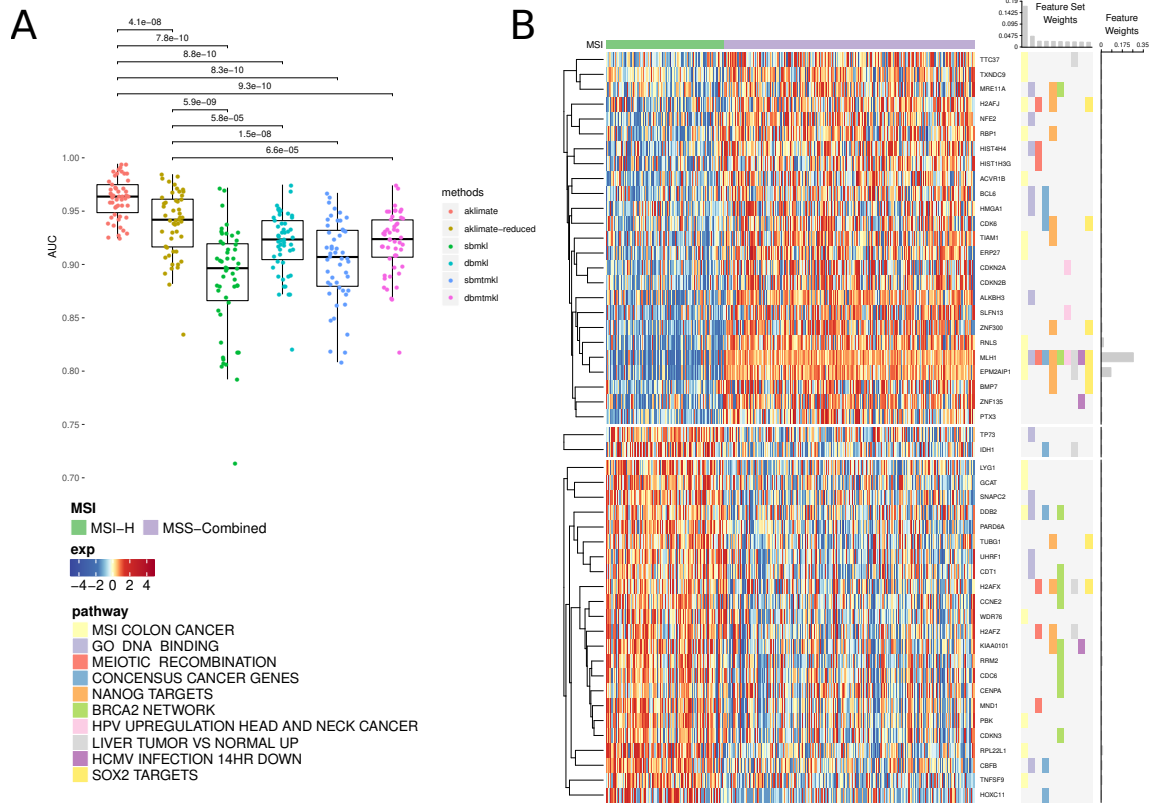


Figure 4.1: AKLIMATE performance on predicting MSI in UCEC TCGA. A) Performance of AKLIMATE and BMKL on classifying MSI-H vs MSI-L+MSS. AUC computed for 50 75%/25% stratified train/test splits. P-values for Wilcoxon signed-rank test pairwise comparisons. Methods: aklimite - aklimite on full collection of feature sets; aklimite-reduced - aklimite with 196 PID pathways; sbmkl - sparse single-task BMKL; dbmkl - dense single-task BMKL; sbmtmkl - sparse multi-task BMKL; dbmtmkl - dense multi-task BMKL. Multi-task BMKL models were trained to simultaneously predict MSI status on UCEC and COADREAD cohorts. B) Top 10 most predictive AKLIMATE feature sets and top 50 most predictive features. Expression of top 50 features (left heatmap); Membership of most predictive features in most predictive feature sets (right heatmap). Features are organized by KNN clustering into 3 groups, followed by hierarchical clustering within each cluster. Feature set model weights scaled to a sum of 1 (barplot, top of right heatmap). Feature model weights scaled to a sum of 1 (Barplot, right of right heatmap). Feature and feature set weights averaged across 50 train/test splits.

pathway constructed by AKLIMATE-reduced represents a poorer approximation to the underlying biological process, as evidenced by its lower AUROC. This is likely due to the fact that only a single PID pathway (out of 196), contains MLH1 ("PID P53 Downstream", 35.2% relative contribution, Fig.4.2), limiting the influence of a key feature on the prediction task.

Interestingly, even though AKLIMATE-reduced used the same feature sets as BMKL, it achieved a statistically significant improvement over all BMKL varieties (Fig.4.1a). This includes both the non-sparse and multi-task BMKL versions, despite the fact that the latter benefited from an entire additional COADREAD data set. In this case, the difference in kernel representation may have contributed to the improved performance of AKLIMATE-reduced (see Section 4.3).

An important aspect of AKLIMATE that highlights the benefits of using prior knowledge feature sets is its relative robustness to false positive findings. Consider AKLIMATE's top two most informative features, MLH1 and EPM2AIP1. EPM2AIP1 is on the opposite strand of MLH1, shares a bi-directional CpG island promoter with it and can be concurrently transcribed [64]. The transcriptional profiles of the two genes are nearly identical (Fig.4.1b) - in the absence of other information (more data types, prior knowledge) it would be extremely difficult to prioritize the "driver" (MLH1) over the "passenger" (EPM2AIP1) using expression data alone. AKLIMATE's feature sets provide just such prior knowledge - while EPM2AIP1 is indeed deemed the second most informative feature, its relative contribution is over three times smaller than that of MLH1. Thus, our method can enrich for biological drivers of MSI with a minimal amount of investigation.

AKLIMATE outperformed other methods on the COADREAD MSI classification task as well. The degree to which it did so was smaller due to the fact that

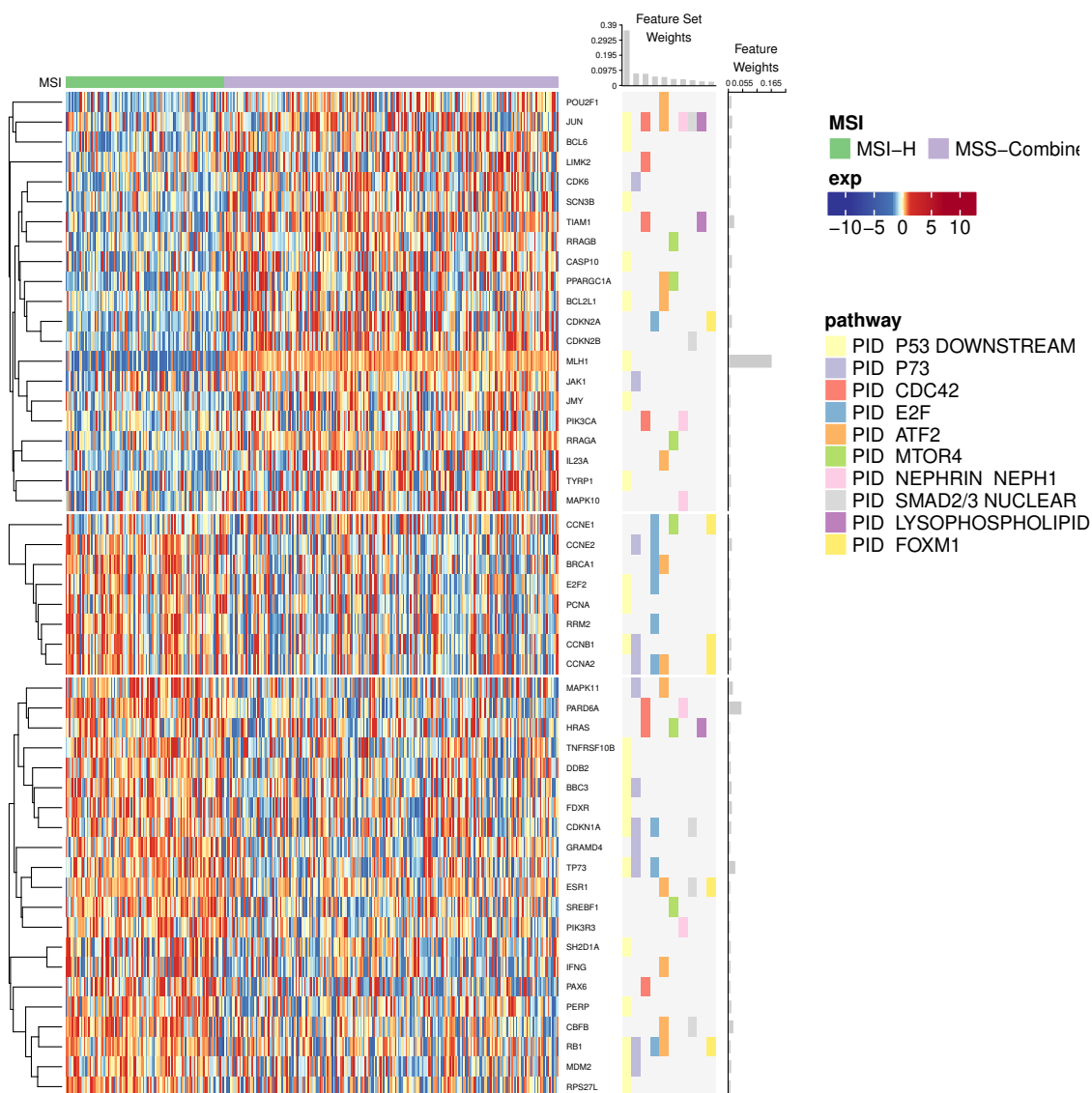


Figure 4.2: AKLIMATE results for the reduced AKLIMATE model (using only 196 PID pathways) on UCEC TCGA cohort. Figure shows top 10 most predictive feature sets and top 50 most predictive features from 50 stratified 75% train/25% test splits. Figure organized as Fig. 4.1.

all classifiers perform well on this problem (Fig.4.3).

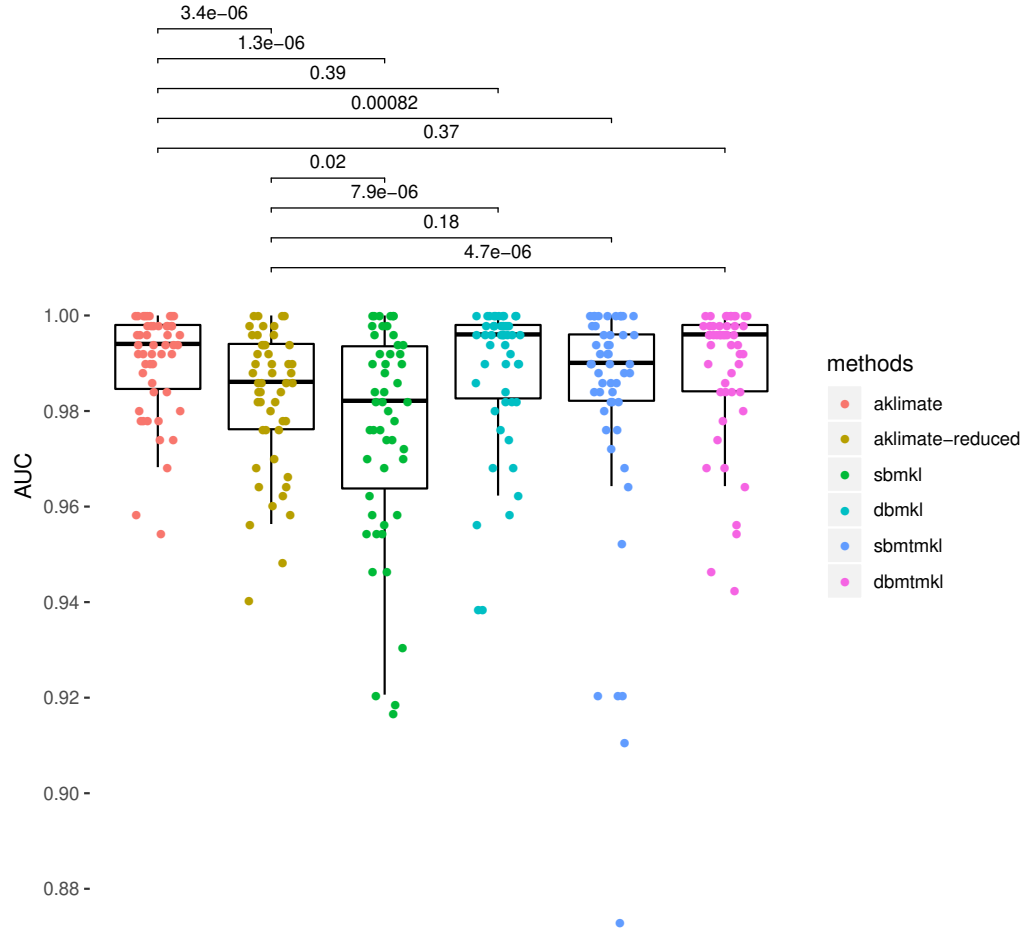


Figure 4.3: AKLIMATE performance on predicting MSI-High vs MSI-Low+MSS in TCGA COADREAD cohort. AUC computed for 50 75%/25% stratified train/test splits. P-values for Wilcoxon signed-rank test pairwise comparisons. Methods as in Fig.4.1.

4.2.2 Breast Cancer Survival

Our next benchmark is predicting survival in the Breast Cancer International Consortium (METABRIC) cohort [38]. This task is significantly more challenging than predicting MSI status, as demonstrated by the DREAM Breast Cancer Challenge [101] and elsewhere [129]. Another big difference between this task and the MSI one is that the METABRIC cohort is annotated with curated clinical data. In fact, the clinical features are quite informative for survival prediction - pre-competition benchmarking by the DREAM Challenge organizers found that models using exclusively clinical features significantly outperformed ones using only genomic features, and did only marginally worse than models in which clinical features were augmented by a subset of molecular features selected through prior domain-specific knowledge [12]. In addition, the best pre-competition clinical feature model was only bettered by the top 5 models in the actual challenge [101].

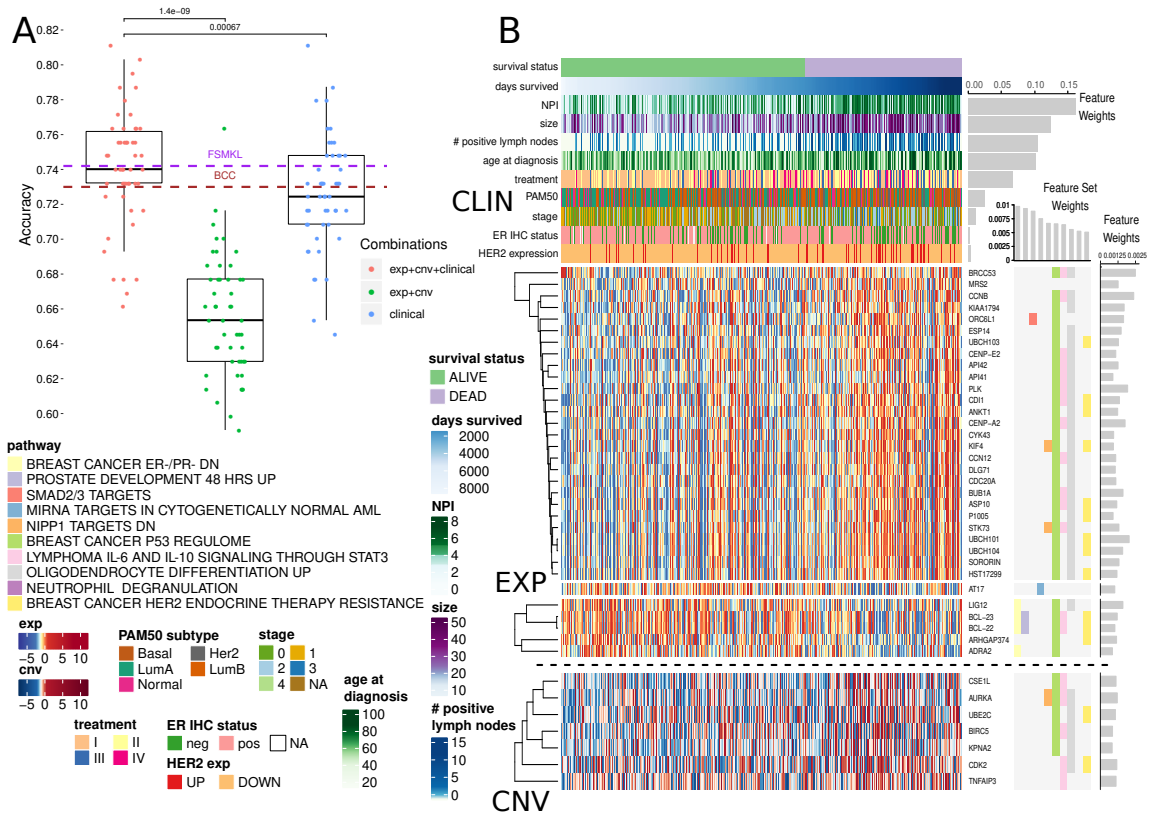
Clearly any model that foregoes the use of clinical data would suffer from inferior accuracy as well as reduced relevance to real-world medical settings. To achieve clinical data integration, AKLIMATE introduces a special category of "global" features - these features are added to every feature set (i.e. the "local" features) prior to the construction of its matching RF. Global features can therefore be interpreted as a uniform conditioning step applied to all AKLIMATE component RFs. In our METABRIC analysis, all clinical features were treated as global.

We compared AKLIMATE to two state-of-the art METABRIC survival predictors. The first one, which we refer to as BCC, is the top-performer in the Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge [101][23] - an en-

semble of Cox regression, gradient boosting regression, and K-nearest neighbors trained on different combinations of clinical variables and molecular-feature derived metagenes. The second one - Feature Selection MKL (FSMKL)[129] - is a pathway-informed extension of SimpleMKL [115] that uses linear and polynomial kernels created from clinical data and molecular features in pathways of the Kyoto Encyclopedia of Genes and Genomes (KEGG) [76]. In FSMKL, pathway features from different data types lead to the construction of separate kernels - in the case of METABRIC, each pathway produces distinct expression and copy number kernels (AKLIMATE, in contrast, learns one kernel matrix from the combined pathway features across all data types). In addition, each clinical feature is treated as a singleton pathway and produces an individual kernel. To make our results directly comparable to FSMKL and BCC as run in [129], we used a subset of the patient cohort ($N = 639$) and a reduced set of clinical variables to match the dataset used in that publication .

We cast the problem as a classification task where we predicted from molecular data and relevant clinical features whether a patient was alive or not at 2000 days. Based on that binarization, there were 387 survivors and 252 non-survivors in our reduced cohort. Similar to our MSI analysis, we performed 50 stratified repeats of 80% train and 20% test partitions; we trained AKLIMATE on each training split and computed accuracy on the respective test samples. To decrease computational time, AKLIMATE’s kernel construction step used 1000 trees instead of the default 2000 (see 4.5.9 for all other hyperparameter settings).

The full AKLIMATE model had higher mean accuracy than BCC (74.1% vs 73.2%) and was on par with FSMKL (74.1% vs 74.2%). This is unsurprising, given that all of these models use clinical variables and correctly prioritize them



as driver features. Furthermore, two of AKLIMATE's main advantages could not be fully utilized under these experimental conditions: first, the benefit of incorporating prior knowledge was reduced because of the relatively low information content of genomic features. Second, the clinical variables did not seem to exhibit complex interaction behavior, which is why modeling them with simpler linear techniques was just as effective. In fact, the most explanatory feature in the full AKLIMATE model (16% relative contribution) was the Nottingham Prognostic

Figure 4.4: AKLIMATE performance on predicting survival at 2000 days in the METABRIC cohort. A) Performance of AKLIMATE under different data type combinations. EXP+CNV - AKLIMATE with genomic features only; clinical - a RF model run with the clinical variables only; EXP+CNV+CLINICAL - AKLIMATE with genomic features as "local" variables and clinical features as "global" ones. FSMKL and BCC dashed lines show mean performances for the two models under 5-fold cross-validation as shown in [129]. B) AKLIMATE results highlighting the top 10 most predictive feature sets and top 50 most predictive features. Figure organized as Fig. 4.1. Clinical variables shown as column annotations; they are included only if among the top 50 most informative features in the model. Clinical variables are ranked from top to bottom by relative predictive contribution. Survival status is a binary variable representing survival at 2000 days (labels) while days survived shows actual duration of survival. Samples sorted by days survived within the two classes. Feature and feature set weights averaged across 50 train/test splits.

Index (NPI)[62] - a linear combination of tumor size, tumor grade, and number of lymph nodes involved. Even with these disadvantages, however, AKLIMATE achieved performance on par with two state-of-the-art methods that were specifically developed on the METABRIC data.

As expected, clinical information proved to be the most influential feature type for predicting outcomes. AKLIMATE models with clinical features alone were significantly more accurate than AKLIMATE models with genomic features alone (p-val=1.9e-08). Even so, the AKLIMATE models using both clinical and genomic features outperformed models with either one alone (p-val=1.4e-09 for full versus genomic, p-val=6.7e-04 for full versus clinical , Fig.4.4a). The mean relative contributions of each feature type in the combined models were 61.6% for clinical, 29.6% for expression, and 8.8% for copy number. The fact that genomic features far outnumbered clinical features (tens of thousands versus 15) underscores the predictive power of the clinical variables and the importance of including them during training.

Of note, while the full AKLIMATE and FSMKL models achieved similar mean accuracies, the importance of individual features and pathways in each model were quite distinct. AKLIMATE heavily favored clinical variables, with 54.5% of the relative explanatory power carried by just five features - NPI, tumor size, lymph node involvement, age and treatment (Fig.4.4b). FSMKL also ranked NPI as the most important clinical feature, but the model considered it less informative than 8 other KEGG pathway-based kernels. Other highly ranked clinical variables included age (11th), histological type(14th), tumor group(34th) and PAM50(40th) [129]. FSMKL’s top ranked pathway was "Intestinal Immune Response," which has some overlap with the "Lymphoma IL-6 and IL-10 signaling through STAT3" signature that AKLIMATE selected as the 7th most informative feature set (Fig.4.4b).

While such distinctions were expected given the different analytical frameworks of the two methods, further analysis is necessary to determine which model aligns more closely with the relevant biology of patient outcomes. Encouragingly, AKLIMATE’s most informative feature sets were enriched for breast cancer progression and response to treatment, with 3 of the top 10 and 8 of the top 20 (Table 4.1) related to these functional groups. For example, the most informative feature set ("BREAST CANCER ER-/PR- DN") represents a signature that is correlated with reduced protein abundance of the estrogen and progesterone hormone receptors [36], while the 9th most informative feature set ("BREAST CANCER HER2 ENDOCRINE THERAPY RESISTANCE") [35] represents transcriptome changes associated with the development of resistance to targeted therapies. Both of these signatures serve as proxies for highly relevant information not available for the METABRIC cohort (protein activity for the progesterone receptor and

therapy resistance). Furthermore, the latest AJCC breast cancer staging manual introduces tumor grade and ER/PR/HER2 receptor status among the key breast cancer classification characteristics, which already include tumor size and lymph node engagement [54]. All of these variables appear as highly informative AKLIMATE model features, either directly or via a proxy genomic signature.

4.2.3 shRNA knockdown viability

We tested AKLIMATE’s ability to integrate multiple data types and solve regression problems on the task of predicting cell line shRNA knockdown profiles. We selected 37 ATARIS [130] consensus profiles for different genes across 216 cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [34] - each consensus profile reflects cell line viability when a particular gene has been knocked down by multiple small hairpin RNAs (shRNAs). We chose these 37 tasks (out of 5711 available consensus profiles) because they had at least 10 cell lines showing strong (>2 sd from mean) viability response to the knockdowns and were in the top quartile by variance of all consensus profiles. We used expression and copy number measurements from CCLE [9] as prediction features. These two data types were augmented by discrete copy number alteration calls made by GISTIC2 [104] (see 4.4) and activities for 447 transcriptional and post-transcriptional regulators inferred by hierarchical VIPER (see Chapter 2). We focused on the 206 cell lines for which we had knockdown profiles, copy number and expression features. Based on results from the DREAM9 gene essentiality prediction challenge [59], we expected this task to be significantly harder than either of our other case studies.

We compared AKLIMATE’s performance to three of the five top performing methods in DREAM9 as well as 3 standard algorithms. The DREAM9 sub-

challenge 1 top performers are briefly described next. Multiple Pathway Learning (MPL) took 5th place (see <https://www.synapse.org/#!Synapse:syn2384331/wiki/64760> for challenge results) - it used elastic-net regularized Multiple Kernel Learning with Gaussian kernels based on feature sets from the Molecular Signature Database (MSIGDB) [93]. MPL and Random Forest Ensemble (MPL-RF) took 2nd place - it computed the prediction averages of a Random Forest classifier and MPL. Both MPL-RF and MPL were extensively discussed in Chapter 3 - we used the same hyperparameters and pathway collections as in the DREAM9 challenge. Kernelized Gaussian Process Regression (GPR) took 3rd place - it used extensive filtering steps to reduce the input feature dimensionality, followed by principal component analysis, and finally Gaussian Process regression with covariance computed from the principal components ([59], see also <https://www.synapse.org/#!Synapse:syn2664852/wiki/68499> for implementation and model description). We downloaded their code from the Synapse URL and ran it with their DREAM9 hyperparameters.

To provide performance baselines, the DREAM9 winners were augmented by standalone Random Forest (RF), Generalized Linear Model (GLM) with a LASSO penalty (GLM-sparse), and GLM with L2 regularization (GLM-dense). RF was run with the *ranger* R package [171] with the following hyperparameters - sampling without replacement with 70% of the samples used for tree construction, minimum node size of 10, 1500 trees and 10% of the features randomly sampled for each node split. GLM-dense and GLM-sparse were run with the *glmnet* R package [51] with the response family set to "gaussian" and the strength of regularization λ learnt through cross-validation. The elastic net tradeoff between the lasso and ridge penalties α was set to $\alpha = 0.8$ (GLM-sparse) and $\alpha = 0.001$ (GLM-dense).

We did not include the nominal first place winner of DREAM9 - an ensemble of four kernel ridge regression models with kernels learned through Kernel Canonical Correlation Analysis and Kernel Target Alignment - because we could not re-run the source code supplied with the challenge submission. We felt this omission is not material as the top 3 methods were declared joint co-winners - their results were shown to not be statistically different from each other but significantly statistically different from the rest of the participants [59]. Furthermore, experiments in Chapter 3 suggested that this method would underperform MPL, MPL-RF and GPR when only high-quality shRNA knockdown profiles are considered.

To reduce correlation across prediction tasks and save computational time, we compare methods on a single stratified train/test split for each ATARIS profile (a different split for each profile) where 67% of the cell lines are used for training and 33% are withheld for testing. Each method was run with its recommended parameters and filtering steps - if no filtering steps were specified, the method used all available features. AKLIMATE’s prediction binarization quantile was set to its default value of $q = 0.05$ (see 4.5.7).

AKLIMATE achieved average Root Mean Squared Error (RMSE) of 1.031 vs 1.047 for GPR, 1.055 for RF, 1.065 for GLM-dense, 1.07 for MPL, 1.071 for GLM-sparse and 1.08 for MPL-RF. The mean RMSE difference was statistically significant in all but one case (Fig.4.5A). AKLIMATE was also the top performer when the number of times an algorithm achieved the best RMSE on an individual prediction task was considered (Fig.4.5B, AKLIMATE retains top spot under different performance metrics as well - Fig.4.6). AKLIMATE performed better than average across nearly all tasks (Fig.4.7); its advantage was particularly pronounced in predicting the essentiality of key regulators (CTNNB1,FOXA1,MDM4,PIK3CA)

or housekeeping genes (PSMC2, PSMC5). As these gene classes are heavily studied and thus well represented in our pathway collections, AKLIMATE’s enhanced accuracy in these tasks may be due to the relatively higher prevalence of relevant feature sets. For example, AKLIMATE’s ability to predict cell viability for MDM4 knockdowns benefits from MDM4’s function as a p53 inhibitor - many of the most informative feature sets in the MDM4 model relate to p53’s regulome and its role in controlling apoptosis, hypoxia and DNA damage control (Fig.4.5C). The top features are also functionally linked to p53. CDKN1A (28.6% relative contribution), a kinase that regulates G1 cell cycle progression, is tightly regulated itself by p53. MDM2 (8% relative contribution) participates in a regulatory feedback loop with p53. ZMAT3 (3.7% relative contribution of expression; 3.5% of cnv) is a zinc finger whose interaction with p53 plays a key role in p53-dependent growth control.

Furthermore, the four p53 features from each data type are all present in the top 50 most informative ones. Individually they carry little signal (relative contribution - 0.5% expression, 0.3% copy number, 0.3% activity, 0.2% cnv gistic) but taken together they clearly implicate p53 as one of the top 10 most informative genes. The ability to capture such multi-omic interactions is one of AKLIMATE’s main strengths. Elucidating the interplay between the different features characterizing p53 would be much harder (if at all possible) in methods that assign separate kernels to each data type, such as BMKL, FSMKL, MPL and MPL-RF. Encouragingly, AKLIMATE dominates the MPL/MPL-RF performance on almost all tasks even though they share the same MKL solver (Fig.4.5A, Fig.4.7).

FOXA1 knockdown prediction is another example of AKLIMATE producing a superior performance on a well-studied gene. FOXA1 dysregulation is an essential

event in breast cancer progression and subtype characterization. Due to breast cancer's status as one of the most prevalent and widely studied types of neoplasm, there is a vast trove of relevant signatures in our feature set compendium. Indeed, 8 of the top 10 feature sets (and 12 of 14 overall) in our AKLIMATE model are directly related to breast cancer experiments under different conditions (Fig. 4.8). As expected, FOXA1 features are the most informative (31.5% relative contribution of expression; 3.4% relative contribution of inferred activity), with AR also among the top 5 most informative genes (2.8% relative contribution of expression).

Out of the 37 shRNA prediction tasks we considered, KRAS was the most obvious example of a well-characterized gene that did not experience discernible accuracy improvement over competing methods (Fig.4.7). We hypothesized that this lack of improvement was caused by the absence of the true biological "driver" from the set of molecular features presented to AKLIMATE. To test that, we added mutation information for 8 key regulators (KRAS, NRAS, PIK3CA, BRAF, PTEN, APC, CTNNB1 and EGFR), 3 of which (KRAS, CTNNB1, PIK3CA) had knockdown profiles among the 37 shRNA prediction tasks. We did not include mutations in our original method comparison because not all cell lines were profiled (165/206 had mutation data) - the 8 regulators were the exception as their mutation profiles covered all 206 cell lines. We included the 8 mutation features as an additional "local" data type instead of the "global" approach we used for clinical variables in the METABRIC case study.

Even though our mutation data type consisted of only 8 features, its addition led to a dramatic improvement in KRAS shRNA prediction accuracy across all metrics (Fig.4.5D, mean RMSE 0.918 ± 0.025 vs 1.02 ± 0.024 ; mean Pearson

0.665 $\pm$ 0.023 vs 0.529 $\pm$ 0.025; mean Spearman 0.57 $\pm$ 0.039 vs 0.453 $\pm$ 0.034). Furthermore, the KRAS mutation feature proved to be by far the most informative (23.5% relative contribution, Fig.4.5E), followed by KRAS copy number (6.9%), KRAS GISTIC (3.3%) and KRAS expression (3.1%). The addition of the KRAS mutation feature was not only crucial in improving the predictive performance of the model, but it also helped prioritize KRAS features from other data types - while KRAS expression, copy number and GISTIC features all appeared among the top 50 most informative in the "no mutation" run, their combined relative contribution was only 3.33% (1.5% copy number, 1% expression, 0.8% GISTIC, Fig.4.9). Such signal amplification is possible because AKLIMATE's RF kernels allow genomic features from different data types to interact directly with one another- this is how complementary genomic patterns (e.g. a composite KRAS pattern drawing from all component data types) can be (correctly) prioritized as most relevant.

Finally, AKLIMATE can examine mutation importance at the amino acid position as well as amino acid substitution level. In the case of KRAS, glycine replacement by either aspartic acid or valine in the 12th amino acid position appeared to have the biggest negative effect on cell viability post-shRNA knockdown (Fig.4.5E). G12 is a well known KRAS mutational hotspot [67] - AKLIMATE's ability to prioritize relevant hotspots could be a key advantage in modeling drug response or recommending treatment strategies.

The same approach did not lead to any predictive benefit in modeling PIK3CA (Fig. 4.10) or CTNNB1 (Fig. 4.11) post-knockdown viability. CTNNB1 has only 9 mutations in the cohort, all of them in different codons. PIK3CA has 30 mutations, with some hotspots - the lack of improvement in this case could be

due to the change in protein sequence not being biologically relevant, the ability of other genomic features to fully capture the mutational signal, or the fact that the "no mutation" models are quite accurate to begin with.

Table 4.1: Most informative feature sets for breast cancer survival prediction in METABRIC data. Weights are averaged over 50 train/test splits. The table lists the top 20 most relevant feature sets, out of 1836 feature sets with a non-zero weight in at least one train/test split. Weights are normalized to sum to 1. Aliases for the top 10 feature sets are included in brackets - they provide a more descriptive representation of biological functionality, based on information gathered from the source publication. The aliases are also used in Fig. 4.4 and throughout the text.

feature sets	weights
GENESIGDB_BREAST_CREIGHTON09_594GENES (BREAST CANCER ER-/PR- DN)	0.0097
SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_UP (PROSTATE DEVELOPMENT 48 HRS UP)	0.0093
KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3 (SMAD2/3 TARGETS)	0.0089
GENESIGDB_LEUKEMIA_MARCUCCI08_696GENES (MIRNA TARGETS IN CYTOGENETICALLY NORMAL AML)	0.0075
NUYTEN_NIPP1_TARGETS_DN (NIPP1 TARGETS DN)	0.0067
GENESIGDB_BREAST_MILLER05_P53 (BREAST CANCER P53 REGULOME)	0.0066
GENESIGDB_LYMPHOMA_LAM08_1502GENES (LYMPHOMA IL-6 AND IL-10 SIGNALING THROUGH STAT3)	0.0064
GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_UP (OLIGODENDROCYTE DIFFERENTIATION UP)	0.0056
NEUTROPHIL_DEGRANULATION_REACTOME (NEUTROPHIL DEGRANULATION)	0.0053
GENESIGDB_BREAST_CREIGHTON08_772GENES (BREAST CANCER HER2 ENDOCRINE THERAPY RESISTANCE)	0.0051
GENESIGDB_BREAST_BARRY10_1022GENES	0.0048
CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_DN	0.0046
MARKEY_RB1_ACUTE_LOF_DN	0.0046
GO_CELL_CYCLE_PHASE	0.0045
GENESIGDB_BREAST_YOSHIHARA10_88GENES	0.0043
GO_MITOTIC_CELL_CYCLE	0.0042
SMID_BREAST_CANCER_BASAL_DN	0.0041
GENESIGDB_OVARIAN_BARANOVA06_907GENES_SERTOLILEYDIG	0.0041
DUTERTRE ESTRADIOL RESPONSE 24HR UP	0.0039
SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP	0.0038

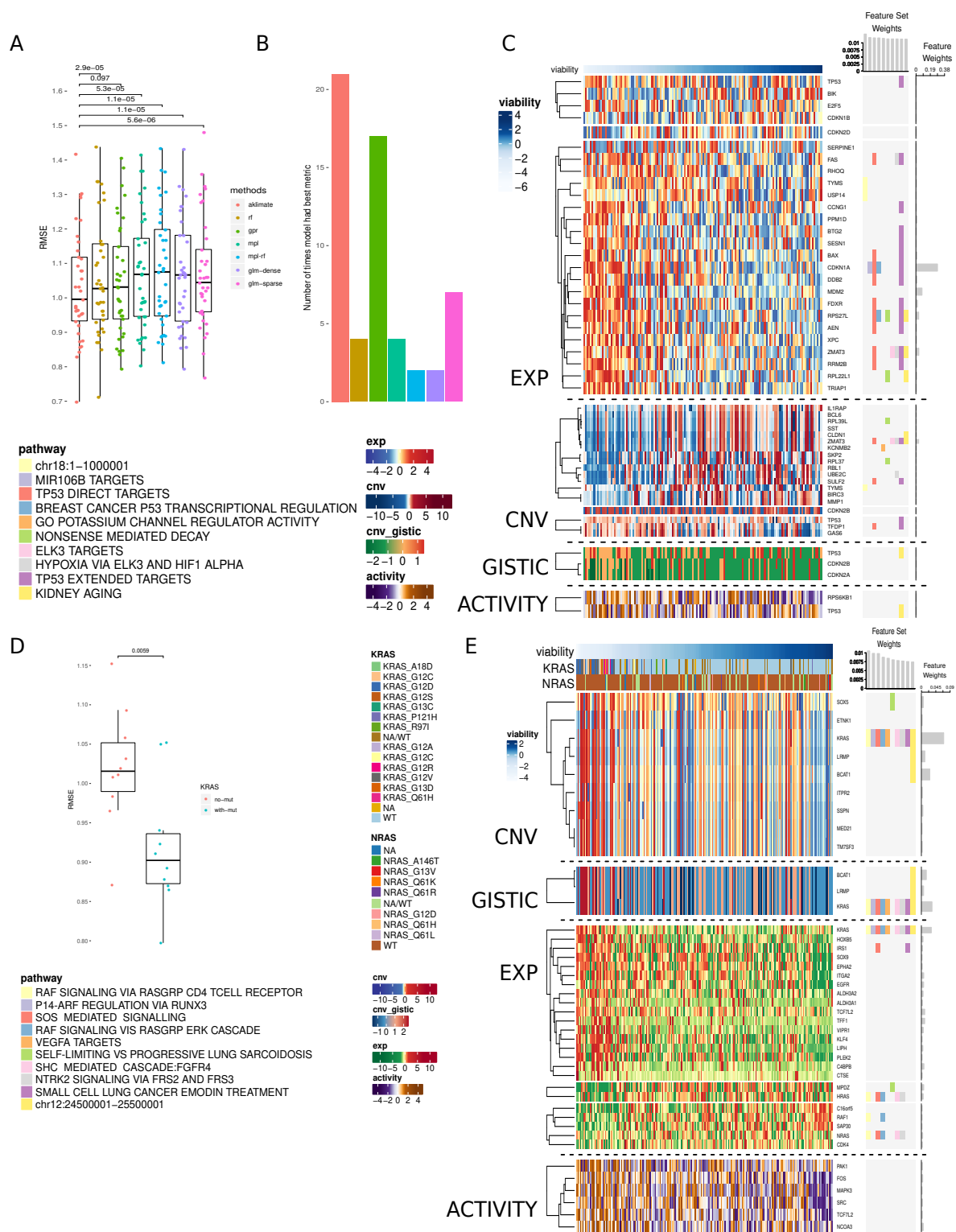


Figure 4.5: Prediction of cell line viability after shRNA gene knockdowns. A) RMSEs of AKLIMATE and competing methods on 37 consensus viability profiles from the Achilles dataset. Methods: Random Forest (RF), Gaussian Process Regression (GPR), Multiple Pathway Learning (MPL), ensemble of MPL and Random Forest (MPL-RF), L2 regularized linear regression (GLM-dense), L1 regularized linear regression (GLM-sparse). B) Number of times an algorithm produced the best RMSE on a prediction task. To prevent small relative RMSE differences from having a biasing effect on the win counts, for each task we consider all algorithms with RMSE within 1% of the min RMSE to be joint winners. For that reason total win counts add up to more than the number of regression tasks. C) AKLIMATE’s top 10 most predictive feature sets and top 50 most predictive features for the task of predicting MDM4 shRNA knockdown viability. Figure organized as Fig. 4.1. D) RMSEs of KRAS AKLIMATE models with and without the use of mutational profiles for 8 key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing. E) AKLIMATE’s top 10 most predictive feature sets and top 50 most predictive features for the task of predicting KRAS shRNA knockdown viability when mutation features are used (feature and feature set weights averaged over 10 train/test splits). Figure organized as Fig. 4.1.

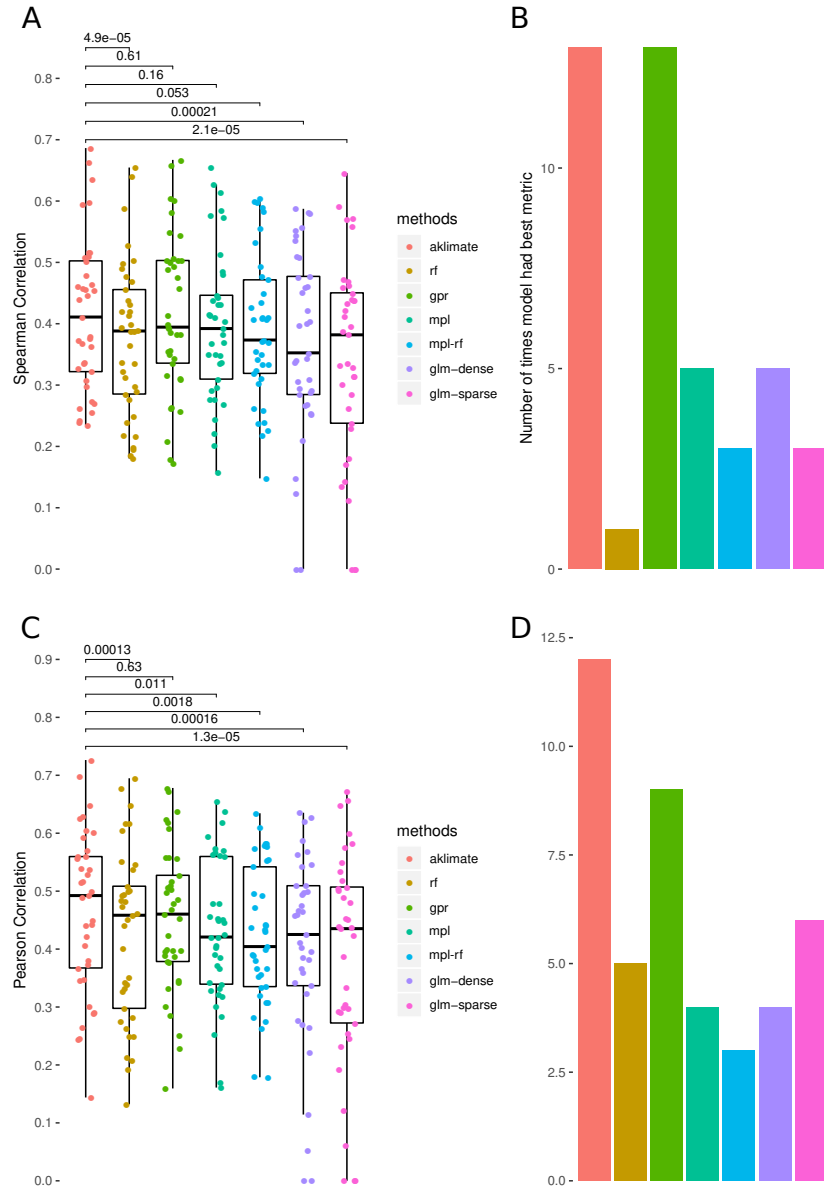


Figure 4.6: Method performance on predicting cell line viability after shRNA gene knockdowns measured by A) Spearman correlation. B) Number of times an algorithm produced the best Spearman correlation on a prediction task. C) Pearson correlation. D) Number of times an algorithm produced the best Pearson correlation on a prediction task.

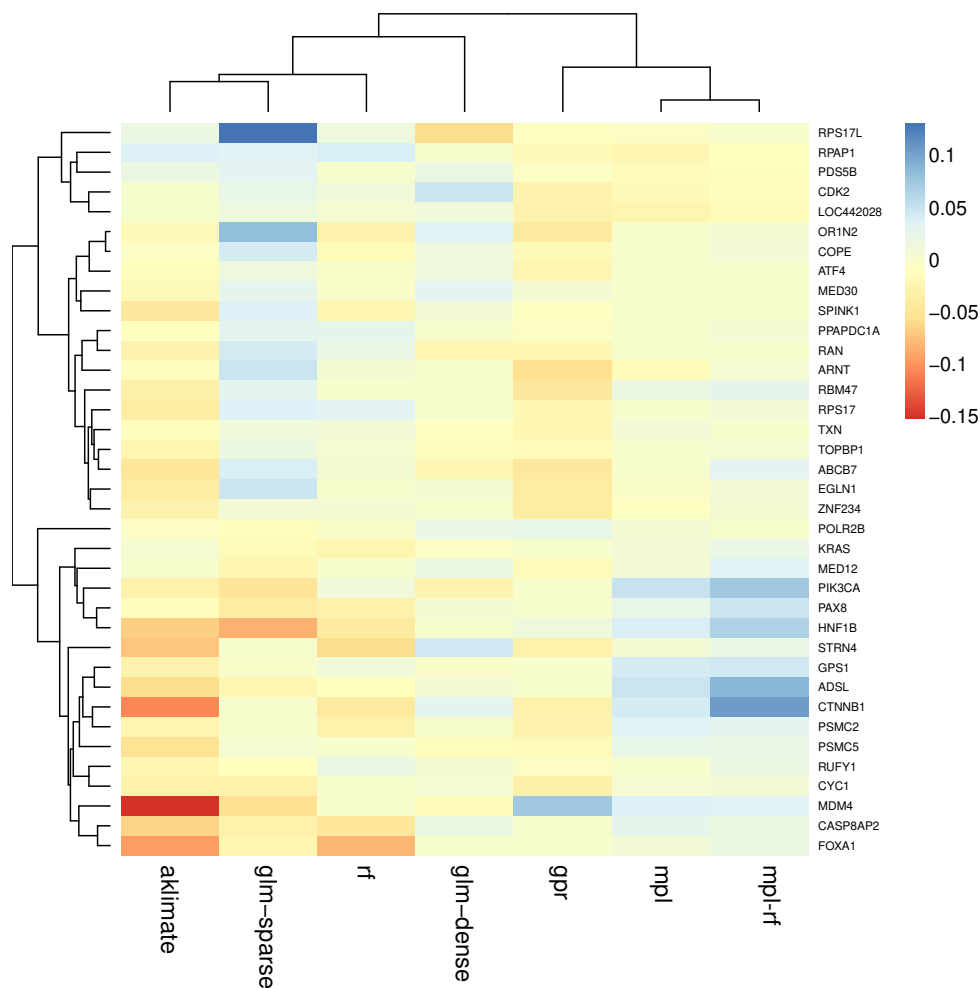


Figure 4.7: RMSE scores for 37 shRNA knockdown prediction tasks. Rows correspond to individual gene knockdowns; columns represent different methods. To highlight differential performance within each task, rows have been centered by subtracting the median. Lower centered scores represent lower RMSE (better performance).

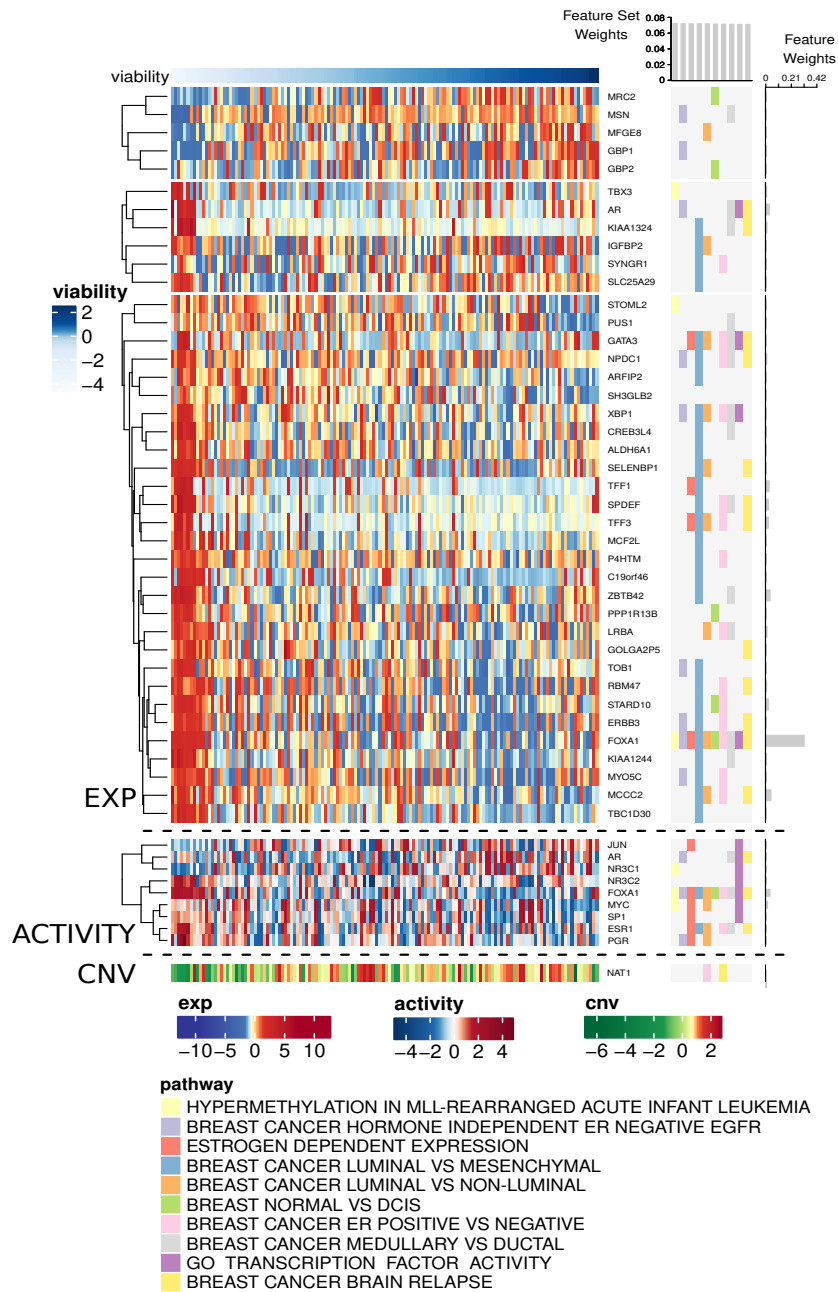


Figure 4.8: AKLIMATE results highlighting the top 10 most informative feature sets and top 50 most informative features for the task of predicting FOXA1 shRNA knockdown viability. Figure organized as Fig. 4.1.

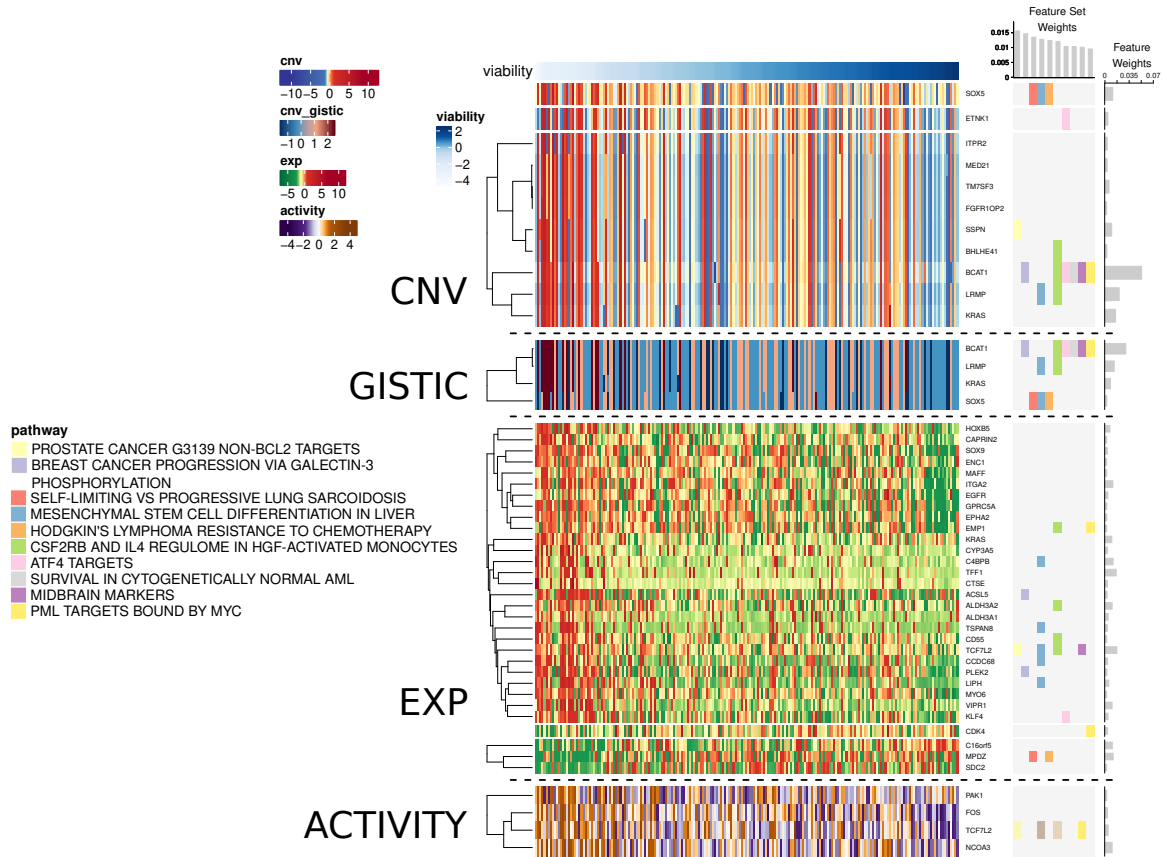


Figure 4.9: AKLIMATE results highlighting the top 10 most predictive feature sets and top 50 most predictive features for the task of predicting KRAS shRNA knockdown viability when no mutation features are used. Feature and feature set weights averaged over 10 train/test splits. Figure organized as Fig. 4.1.

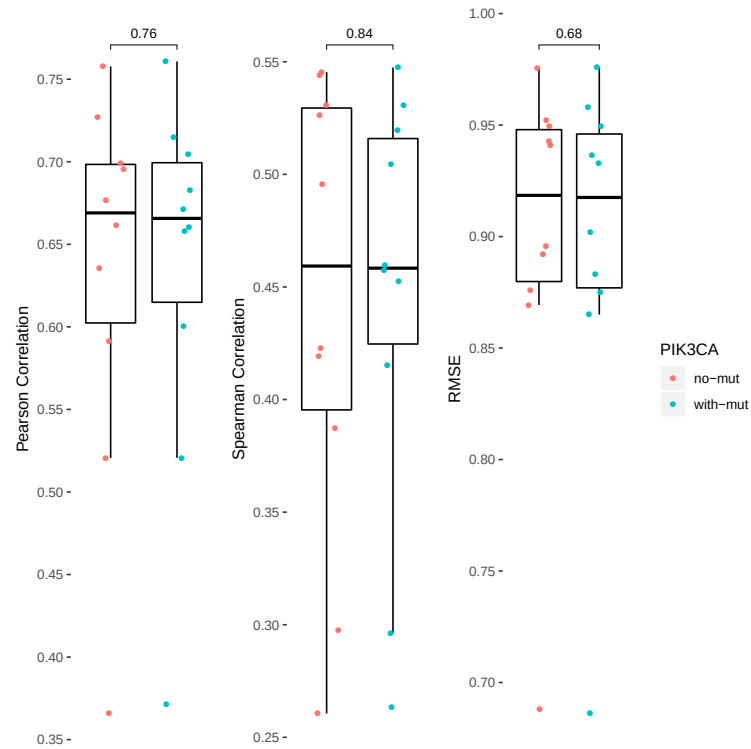


Figure 4.10: Metrics for PIK3CA AKLIMATE models with and without the use of mutational profiles for 8 key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing.

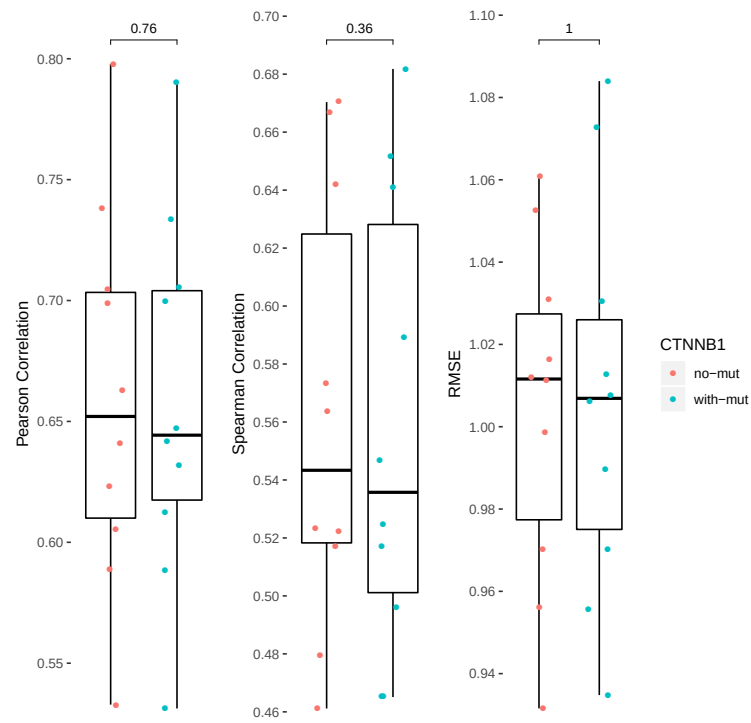


Figure 4.11: Metrics for CTNNB1 AKLIMATE models with and without the use of mutational profiles for 8 key regulators. Results shown for 10 matched stratified train/test splits where 80% of the cohort is used for training and 20% for testing.

4.3 Discussion

Recent surveys of cancer genome landscapes have shown that alterations of a particular pathway can involve many different genes and many different kinds of disruptions - for example, RB1 mutation, RB1 methylation, or CDKN2A deletion could all lead to cell proliferation [42]. Consequently, combining data at the level of pathways may benefit machine-learning applications in the cancer genomics setting. However, data platform diversity often prohibits such integration - variables can be of different scales (e.g. copy number vs gene expression) or different types (continuous DNA methylation, binary mutation calls, ordinal inferred copy number estimates). AKLIMATE's early integration approach is a potential solution to capturing complementary pathway-level information spread across data modalities - all data types are considered, and potentially used, when constructing every pathway kernel. AKLIMATE does so by building supervised tree-based empirical kernel functions that optimally align the training labels with each pathway-specific set of multimodal data. In contrast, FSMKL (and other MKL approaches with unsupervised kernel construction) integrate pathway knowledge and multiple data types by computing segregated data type-specific kernels for each pathway and letting the linear combination "meta-kernel" determine each data type's contribution. This may result in suboptimal solutions - features that belong to the same pathway but in different data types can now only interact with each other on the kernel level and not individually. AKLIMATE's approach creates a much richer interaction model that is flexible enough to capture same-gene, cross-gene, and cross-data type interactions.

Another limitation of current pathway-informed kernel learning methods is that a single informative feature can go undetected if only present in large path-

ways - if all member features contribute equally to kernel construction, the importance of the relevant feature is obscured by the non-relevant majority. In contrast, AKLIMATE's RF-based kernels effectively allow individual features to influence the model. This advantage is illustrated by the improvement in AKLIMATE performance over BMKL on the MSI prediction task. BMKL's Gaussian kernels treat each feature in a feature set as equally important in the computation of the respective kernel matrix. As MLH1 appears only once in the PID pathway set, its contribution is masked by less informative features. On the other hand, AKLIMATE's RF kernels, due to the supervised manner in which they are constructed, inherit RF's ability to prioritize features based on their importance to the classification task - informative features are by definition overrepresented among tree node splitting variables. As all components of the RF kernel are derived from characteristics of the RF trees, informative features exercise proportionately higher influence over the RF kernel construction. Therefore, if only a small subset of a pathway's features are truly relevant, they can be clearly distinguished from (thousands of) non-relevant ones. For example, both AKLIMATE and AKLIMATE-reduced picked MLH1 as the most informative feature (26.7% and 16.4% relative contribution respectively), with a steep decline in the importance of the next best feature (AKLIMATE - EPM2AIP1, 8% relative contribution; AKLIMATE-reduced - PARD6A, 4.8% relative contribution).

The METABRIC case study highlights another key advantage of AKLIMATE - a straightforward way to accommodate variables that do not readily map to genome-based feature sets (e.g. clinical data). In cases such as METABRIC, where clinical features provide most of the predictive power, the most salient question becomes what genomic features can provide additional orthogonal information

given that the clinical data are already incorporated into the model. Posing the problem as a conditional relationship between clinical and molecular data closely reflects the practical situation in hospital settings where clinical information is usually the first thing available to a treating physician. AKLIMATE and FSMKL illustrate two different ways of incorporating such information. FSMKL treats each clinical variable as a feature set of size one and constructs a kernel for each of them individually. This approach is viable, but quite restrictive in how it models clinical variable interactions. While AKLIMATE can accomodate such a setup, it also permits a more complex representation of the way features interact - each clinical feature is of a special "global" type that gets included in every feature set in addition to the features "local" to it. The "global"-"local" feature hierarchy allows maximum flexibility in modeling interactions among clinical variables and between clinical variables and genomic features. Such a hierarchy is necessary when features work on different biological scales - for example, tumor grade is an organ level characteristic that captures a snapshot of the behavior of millions of cells and is therefore likely to be vastly more informative than the copy number status of an individual gene.

An important aspect of AKLIMATE's use of prior knowledge is its ability to identify relevant features even in cases where many confounders exhibit high collinearity. This problem is similar to the one encountered in genome-wide association studies where an allele conferring a phenotype of interest can exist in a large haplotype block containing the alleles of many other genes. In such situations, prior knowledge often helps researchers select the true causal variant among the many false positive ones. In the MSI prediction task, AKLIMATE was able to identify MLH1 among several other correlated features based solely on gene

expression data. AKLIMATE’s robustness to collinearity is not limited to expression features. Importantly, AKLIMATE can prioritize relevant genes even if they are subject to large-scale copy number events and thus have almost identical copy number profiles with many other genes. This can be observed in AKLIMATE results for the MDM4 knockdown prediction task (Fig.4.5C) as well as the consensus AKLIMATE feature importance for KRAS predictions with mutation data (Fig.4.5E). In the former, there is a clear large scale CN event that involves 7 of the top 50 most predictive genes, but ZMAT3 is given by far the highest weight because of the biological prior of the feature sets. Similarly, in the latter 9 CN features have very similar profiles, but the KRAS one is prioritized as the most important. The KRAS GISTIC feature is also favored among a group of genes affected by large-scale events - a strong indicator that the robustness to collinearity extends beyond continuous features to categorical ones as well.

Making use of prior knowledge, such as the results of past experiments, is often a key component to the success of machine-learning applications in genomics analysis [12] [147]. AKLIMATE uses a biologically-motivated prior distribution on the feature space - as demonstrated, this approach often outperforms methods that use a uniform prior over input features. AKLIMATE updates its prior information in a data driven manner - therefore, the feature set compendium does not need to be specifically tailored to the problem at hand, avoiding the need for feature set membership filtering heuristics. When high quality relevant prior experiments are available, AKLIMATE tends to perform better. For example, the most important feature set for MSI prediction in endometrial cancer was a previously published signature characterizing MSI in colon cancer (Fig.4.1B). From that perspective, AKLIMATE can be thought of as a framework for discovering

the past experiments whose results are most relevant to the interpretation of a new dataset.

Alternatively, AKLIMATE can be considered a data type prioritizer. By aggregating the weights of features within a data type, AKLIMATE can implicitly measure data type relevance to an experiment. For example, while expression is generally most important in predicting shRNA knockdowns (mean importance 71.6% across tasks), there are cases where copy number is more informative - PSMC2(61.7%), RPAP1(55%) and CASP8AP2(47%) (Fig.4.12). Furthermore, inferred protein activity varies tremendously in terms of its contribution - from $< 1\%$ in predicting PSMC2 knockdowns to 27.7% for STRN4 (Fig.4.12). AKLIMATE’s ability to zero in on information-rich data types could help in designing targeted future experiments.

From a stacked learning perspective, AKLIMATE implements level-one data augmentation by including not just a base learner’s CV-based predictions, but also its intrinsic properties that might convey additional predictive information (in the case of AKLIMATE base RFs - how often data points end up in the same leaf node and whether they tend to end up in early-split or late-split leaves). This augmentation leads to improved performance - if we use AKLIMATE’s *bestRFs* as the base learner set, AKLIMATE outperforms with statistical significance the ensemble that averages base learner predictions as well as the best individual base learner across all three case studies (Fig. 4.13). In addition, it does better than a standard Super Learner (regularized linear regression meta learner) although without achieving statistical significance on some of the data sets (Fig. 4.13). This clearly indicates that the additional information encoded in components K_2 and K_3 of AKLIMATE’s *RF* kernels is relevant and helpful. What is more, it

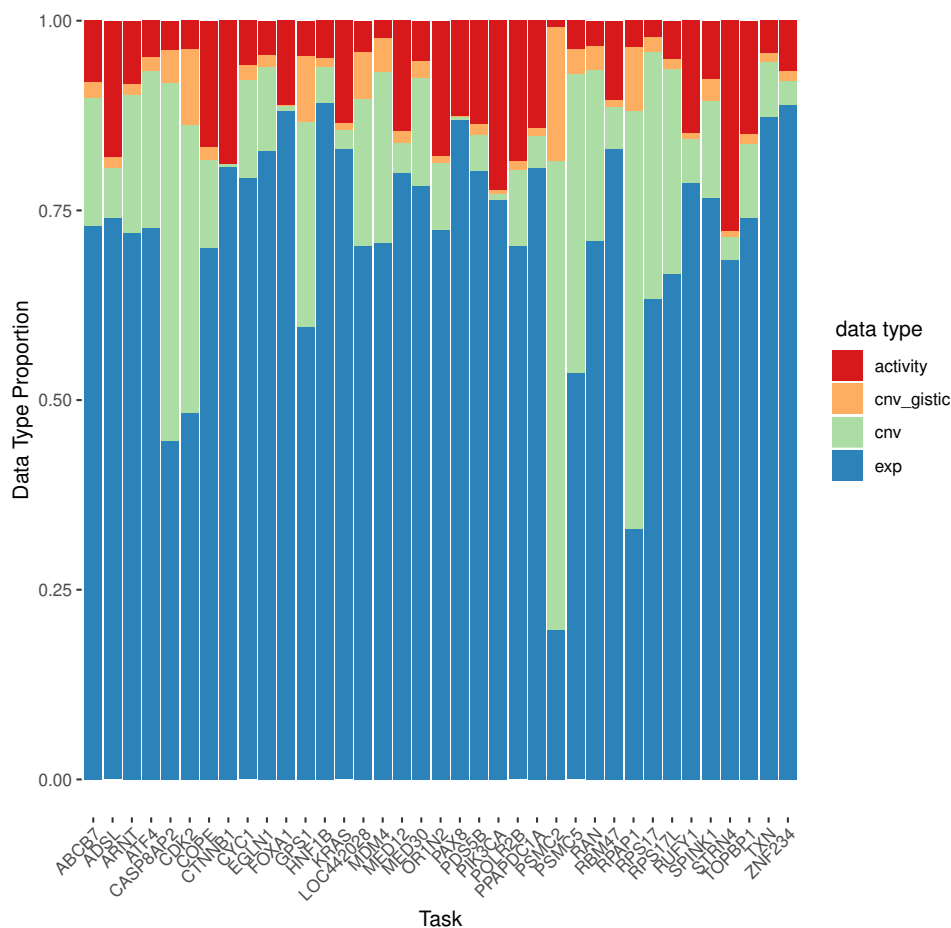


Figure 4.12: Data type importance proportions for 37 shRNA knockdown prediction tasks. Each column corresponds to the relative contributions (normalized to a sum of 1) of the different input data types for a particular prediction task. Relative contributions are computed by summing up the model-assigned feature importance scores within individual data types.

highlights how intrinsic properties of a base learner can be propagated into a more accurate meta learner. While AKLIMATE provides a blueprint for tree-based algorithms, using other base learners might yield even better results.

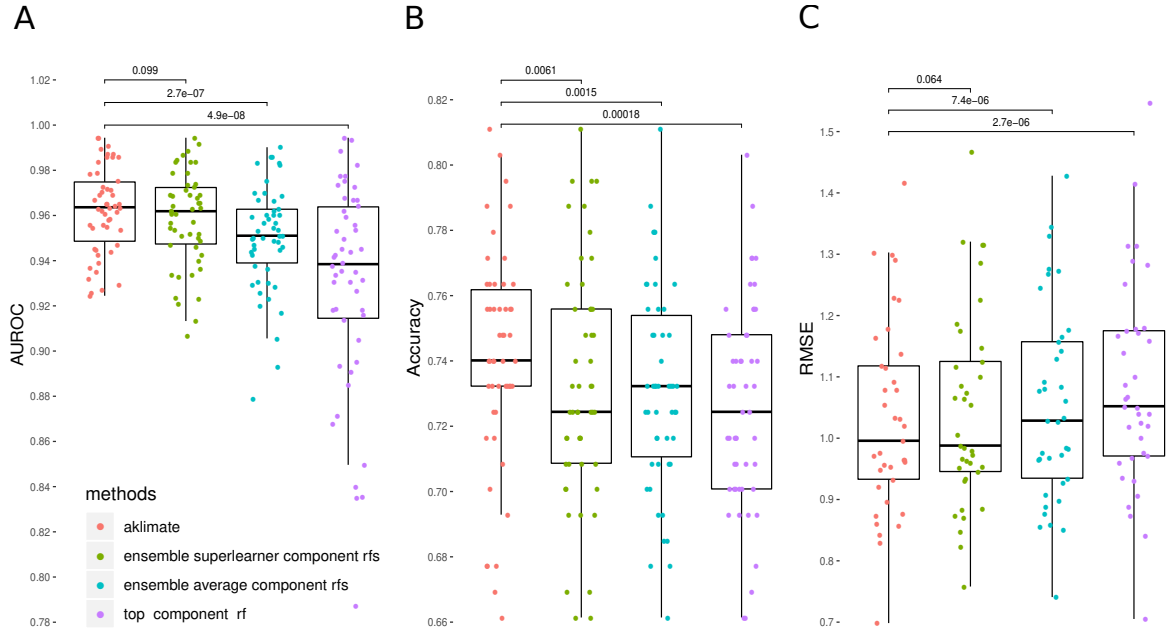


Figure 4.13: Comparison of AKLIMATE versus different ways of combining AKLIMATE component RFs. Ensemble superlearner component RFs - learning a regularized linear regression of predictions from component RFs; ensemble average component RFs - taking the average of the predictions of component RFs; top component RF - only using predictions from the top ranked RF. A) UCEC TCGA MSI prediction. B) METABRIC Breast Cancer survival prediction. C) Achilles shRNA knockdown prediction. P-values for Wilcoxon signed-rank test pairwise comparisons.

AKLIMATE requires minimal feature pre-processing, can query tens of thousands of feature sets, and its main steps are embarrassingly parallel. It performs as well as, or better than, state-of-the-art algorithms in a variety of prediction tasks, with the added benefit of easily interpretable integrated feature set analysis. AKLIMATE can natively handle continuous, binary, categorical, ordinal and

count data, and its feature sets can be trivially extended with features related to current members - e.g. gene sets can be augmented by appropriate methylation probe status, promoter mutations, splice variant proportions and many more. Furthermore, with slightly more effort AKLIMATE can incorporate prior importance scores for features within a feature set as well as structured prior relationships such as known feature-feature interactions or dependencies. Finally, even though all case studies came from the field of bioinformatics, AKLIMATE can be applied to any problem with prior knowledge of problem-relevant feature groups and/or heterogeneous data types, particularly problems where the feature sets extend across data types.

4.4 Data Acquisition and AKLIMATE pre-processing

Data for all three case studies was processed in the following manner:

1. Expression data was filtered based on the mean and variance of genes across samples - any gene whose mean or variance fell in the bottom 20% of the respective distribution was discarded.
2. Copy number data, if available, was filtered similarly to expression but with the cutoff set at 50%. In case GISTIC2 [104] discretized gene-level copy number calls were computed, they were filtered the same way.
3. hVIPER-generated regulator activity scores were not filtered as there were a small number of such features to begin with (447, see Chapter 2 for hVIPER description and derivation).

In addition, filtered data for the two classification tasks (MSI and METABRIC

survival prediction) was discretized by computing the quintiles of the distribution of each molecular feature and binning each quintile into a separate category. This discretization was done to speed up computation in the larger cohorts. Finally, non-ordered categorical data (e.g. METABRIC clinical variables) were one-hot encoded prior to AKLIMATE use.

4.4.1 Microsatellite Instability

Data for the COADREAD and UCEC TCGA cohorts were downloaded from the Synapse copy of the PANCAN12 TCGA cohort [65] (synapse object id syn300013, <https://www.synapse.org/#!/Synapse:syn300013/wiki/70804>). The UCEC expression data (syn1446289) was log transformed and filtered as previously described. The filtered data set contained 13,424 expression features. The MSI status for UCEC patients was extracted from UCEC clinical data (syn1446167).

Similarly, the COADREAD expression dataset was created by joining the COAD (syn1446197) and READ (syn1446276) PANCAN12 cohorts, log transforming the combined matrix and applying previously described filtering. The filtered data set contained 14,036 features. MSI status was downloaded from firebrowse.org and matched to the whitelisted samples for the joint COAD-READ PANCAN12 expression set.

4.4.2 METABRIC Survival

Expression (Illumina HT12 array), copy number (Affymetrix SNP 6.0) and clinical data for the METABRIC cohort[38] were downloaded from <https://www.synapse.org> (Synapse ID syn1688369). Expression and copy number data were processed as described in the marker paper [38]. Expression data used Illumina

HT12V3 probe identifiers while copy number data had Entrez gene features. Since our pathway compendium is HGNC-based, we translated each HGNC gene set to all Illumina probes and Entrez gene ids matching any of its members. We used the *IlluminaHumanv4.db* Bioconductor package for the Illumina-HGNC map and the *org.Hs.eg.db* package for the Entrez-HGNC one.

To match the analysis in [129], we restricted the METABRIC cohort to 639 patients (list obtained in personal communication with authors). For the same reason, we did not use the full set of clinical information available, but limited it to variables used in [129], namely:

1. Age at diagnosis
2. Tumor size
3. Tumor grade
4. Tumor stage
5. Number of positive lymph nodes
6. Histological type
7. Estrogen receptor IHC status and expression-based status
8. HER2 IHC status, SNP6 status, and expression-based status
9. Nottingham Prognostic Index
10. PAM50-based breast cancer subtype
11. Cellularity
12. Composite treatment status

The last clinical variable is not present in the METABRIC clinical file, but it integrates aspects of other clinical variables in a manner described in [129].

Expression and copy number data sets were filtered as previously described, with mean and variance calculations based on the reduced rather than the full cohort. The combined filtered feature set contained 20,022 expression, 8608 copy number and 15 clinical features.

4.4.3 shRNA Knockdown Profiles

Achilles 2.4.3 shRNA knockdown profiles were downloaded from <https://depmap.org>. The data release contained ATARIS [130] gene-level profiles for 5711 genes across 216 CCLE cell lines. ATARIS profiles were computed by aggregating the profiles of multiple shRNAs targeting an individual gene. ATARIS was run with a threshold of $p = 0.05$ on the samples and shRNAs that passed QC inspection (see online QC manifest of the Achilles 2.4.3 data). The mutation profiles of the 8 regulators considered in 4.2.3 (KRAS, NRAS, PIK3CA, BRAF, PTEN, APC, CTNNB1 and EGFR) were extracted from the sample annotation file for the Achilles 2.4.3 data release.

Matching expression and copy number characterizations of individual cell lines were downloaded from the Cancer Cell Line Encyclopedia (CCLE, <https://portals.broadinstitute.org/ccle/data>). Expression was measured by Affymetrix U133 Plus 2.0 array, aggregated via Robust Multi-array Average and quantile normalized (see 2012 expression data release on CCLE website). Copy number was evaluated by Affymetrix SNP 6.0 arrays and segmentation of the normalized log2 probe ratios via Circular Binary Segmentation (see 2012 copy number release on CCLE website). GISTIC2 version 2.0.22 was run on the copy number data

with default parameters (amplification and deletion thresholds of 0.1, broad event threshold of 0.7, enabled arm level peel-off events, and gene collapsing set to extreme). hVIPER was run on the expression data as described in Chapter 2.

Expression, copy number and GISTIC data sets were filtered as previously described. The combined filtered feature set contained 13,652 expression, 10,086 copy number, 9,557 GISTIC and 447 VIPER activity features.

4.4.4 Gene Set Collections

As AKLIMATE can feasibly query several times more gene sets than MPL (see 4.5.7), we developed a much more extensive gene set compendium for use with it. We again included the C2 and C5 collections of MSigDB [139], but we augmented them with GeneSigDB [37] (a curated collection of published signatures) and PathwayCommons [21] (a database of databases covering the spectrum of metabolic, molecular, signaling, regulatory and genetic interactions). We excluded gene sets from the Small Molecular Pathway Database [73] (SMPDB, part of Pathway Commons) because of its highly redundant and small-sized signatures. We also removed the "canonical" pathways C2_CP MSigDB subcollection due to its high degree of overlap with the much more extensive PathwayCommons resource. Furthermore, we added gene sets related to chromosomal location - they were derived by passing TCGA LIHC segmented copy number data through the *CNRegions* function of the *iClusterPlus* [105] R package, with $\epsilon = 0.0025$. Finally, we removed all sets with more than 1000 members. The final compendium consists of 17,273 gene sets with median size of 30 (min size of 1, max size of 991).

4.5 Methods

AKLIMATE is a stacked learning framework in which each base learner encapsulates the relationship between the training labels and a biologically relevant concept mapped to a group of multi-modal features (a feature set, Fig.4.14). The base learners are RFs, which allows the concurrent use of features with different underlying distributions without the need for cross-data type normalization. The feature sets can come from many diverse sources - e.g. pathway compendiums, chromosomal regions, drug treatment differential signatures, siRNA knockdown experiments or disease progression studies. Furthermore, even though such feature sets normally consist of member genes, they are easily amenable to a more comprehensive membership - for example, if such input data are available, gene sets can be extended by features corresponding to mutational hotspots, different splice forms, or nearby methylation probes. Base learners are agnostic to any features not included in their associated feature set - as a consequence, each one presents a distinct, if correlated, view of the data. Each base learner produces a pairwise similarity measure on the training samples (an RF kernel, see 4.5.4) that captures the contributions of all modes of data mapping to the relevant feature set. The stacked learning step finds the optimal combination of such similarity matrices - it can be interpreted as the kernel associated with the most predictive integrated meta-feature set derived from all interrogated feature sets.

The key contributions of AKLIMATE are:

1. The introduction of an integrative empirical kernel function that combines similarity in predicted labels with proximity in the space of RF trees (RF kernel).

2. The extension of kernel learning to a stacking framework. To our knowledge, AKLIMATE is the first stacked learning formulation to incorporate base kernels.

In the next sections we give a detailed description of each AKLIMATE component.

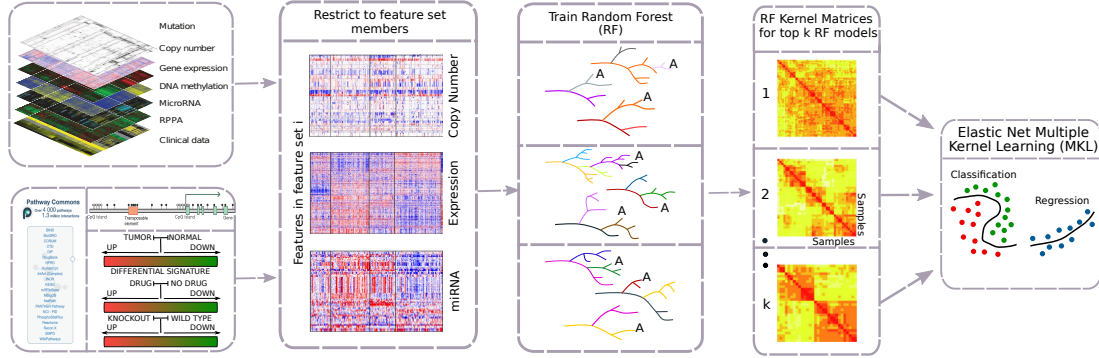


Figure 4.14: Overview of AKLIMATE. AKLIMATE takes as inputs multiple data types and a collection of feature set groups. AKLIMATE first trains RF learners, one for each feature set, with all available multi-omic features that map to the set in question. The RFs are then ranked by their OOB performance and the top K are converted to RF kernels. Finally, RF kernels are used as input in an elastic net MKL to produce the final predictions. Elastic net hyperparameters are learnt via cross-validation.

4.5.1 Random Forest

An RF learner is an ensemble of decision trees, each of which operates on a perturbed version of the training set. The perturbation is achieved by two randomization techniques - on one hand, each tree trains on bootstrapped data generated by drawing samples with replacement from the training set. On the other hand, at each node in a decision tree only a subset of all features are considered as candidates for the next split. Such randomization is motivated by the

desire to decorrelate the predictions of individual decision trees - averaging predictions of uncorrelated models tends to reduce the variance component in the bias-variance decomposition of prediction error. Since decision trees are generally low-bias high-variance predictors, in most applications the lower variance of the RF ensemble far outweighs the accompanying increase in bias, leading to lower overall prediction error (for an in-depth examination, see [94]).

4.5.2 Kernel Learning

Kernelization is a widely-used technique that allows linear discriminant methods to be applied to learning problems with non-linear decision boundaries [125]. It relies on the "kernel trick" - a transformation of the input space to a different, potentially infinite dimensional, feature space where the train set samples are linearly separable. The utility of the "kernel trick" lies in the fact that the feature map between the input space and the feature space does not need to be explicitly stated - if an algorithm can be formulated in terms of dot product calculations, we only need to know the functional form of the feature space dot product in terms of the input space variables. A kernel function represents such a generalized dot product [125]. Furthermore, the Representer Theorem [80][125] guarantees computational tractability since the optimal solution to a large collection of optimization problems is guaranteed to only depend on kernel function evaluations on the (finite-dimensional) training data. More formally, for an arbitrary loss function $L(f(x_1) \dots f(x_N))$, the minimizer f^* of the regularized risk function

$$R(f) = L(f(x_1) \dots f(x_N)) + \Omega(\|f\|_H^2) \quad (4.1)$$

can be expressed as

$$f^* = \sum_{i=1}^N \alpha_i k(x_i, \cdot), \alpha_i \geq 0 \quad (4.2)$$

provided that the regularization term $\Omega(\|f\|_H^2)$ is a monotonically increasing function.

The kernel trick is valid under mild conditions - the kernel needs to be positive-definite, ensuring the existence of an associated Reproducing Kernel Hilbert Space [125][4]. Methods such as Support Vector Machines, Gaussian Processes, Fisher Discriminant Analysis [125] and Kernel PCA [126] have all been extended to non-linear settings via kernelization.

4.5.3 Multiple Kernel Learning

Kernel learning can lead to large improvements in accuracy, provided that an appropriate kernel is selected. That choice, however, is often not obvious - kernel functions might not be easily prioritizable, they might have hyperparameters that need to be tuned or the training data might consist of heterogeneous features that do not natively conform to a single feature map. MKL addresses the problem of kernel selection by searching for a composite kernel that is a data-driven optimal combination of candidate kernels [7][88][58][82]. Linear combinations of kernels of the form $K(\alpha) = \sum_{i=1}^M \alpha_i K_i$ with either a conical ($\forall i, \alpha_i \geq 0$) or convex ($\forall i, \alpha_i \geq 0$ and $\sum_{i=1}^M \alpha_i = 1$) constraint are by far the most commonly used target kernel representations because of their many desirable properties (although algorithms that allow non-linear combinations do exist - see [58]). Some important advantages are:

1. Conical/convex combinations of positive definite kernels are positive definite.
2. The composite kernel is associated with a feature space that is the concatenation of all component kernel feature spaces.
3. The Representer Theorem is readily extendable to the the conical combination case - the optimal solution f^* takes the form $f^* = \sum_{m=1}^M \sum_{i=1}^N k_m(x, x_i) \alpha_{m,i}$ [145][125].

MKL algorithms admit different forms of regularization depending on what norm of the kernel weights is chosen. In our work we use an elastic-net [176] regularizer on the norm of the individual kernel weights of the form $\lambda_1 \sum_{i=1}^M \|\alpha_m\|_{K_m} + \lambda_2 \sum_{i=1}^M \|\alpha_m\|_{K_m}^2$, where $\forall \lambda \geq 0$, $\alpha_m = (\alpha_{m,1} \dots \alpha_{m,N})^T$, and $\|\alpha_m\|_{K_m} = \sqrt{\alpha_m K_m \alpha_m}$ is the definition of a kernel norm as in [145] and [144]. We chose elastic net regularization because of its flexibility - with appropriately tuned λ 's we can capture sparse ($\lambda_1 \gg \lambda_2$, few non-zero weight norms) and dense ($\lambda_1 < \lambda_2$, many non-zero weight norms) kernel weight solutions. Of note, such a regularization also permits a uniform kernel weight solution ($\lambda_1 = 0$, $\lambda_2 > 0$).

The explicit form of our optimization problem becomes

$$\min_{\alpha \in \mathcal{R}^{NM}, b \in \mathcal{R}} \sum_{i=1}^N \mathcal{L}(y_i, \sum_{m=1}^M \sum_{j=1}^N k_m(x_i, x_j) \alpha_{m,j} + b) + \lambda_1 \sum_{i=1}^M \|\alpha_m\|_{K_m} + \lambda_2 \sum_{i=1}^M \|\alpha_m\|_{K_m}^2 \quad (4.3)$$

as described in [145]. We use their algorithm (SpicyMKL) to solve (4.3).

The optimal kernel weights are recovered by

$$w_m = \begin{cases} 0 & (\|\alpha_m^*\|_{K_m} = 0), \\ \frac{\|\alpha_m^*\|_{K_m}}{\lambda_1 + \lambda_2 \|\alpha_m^*\|_{K_m}} & (\text{otherwise}) \end{cases} \quad (4.4)$$

where α^* is the optimal solution of (4.3) [145]. The model weights are also re-scaled to satisfy $\sum_{i=1}^M w_i = 1$.

4.5.4 RF Kernel

The most common approach to kernel construction is to specify an explicit closed form for the kernel function, such as a polynomial or a radial basis one. Kernel functions that encode more complex relationships between training objects also exist, particularly if the objects can be defined in a recursive manner (ANOVA kernels, string kernels, graph kernels, see [131]). However, each of these kernel functions represents a data dependency model chosen prior to kernel construction. Such an approach can suffer from lack of robustness particularly when it is not obvious what the appropriate data dependency model is. Learning that dependency structure from training data is a logical solution, particularly if the procedure is generalizable and resistant to overfitting. We propose the use of random forests as tools for discovering the empirical dependency model and for converting it to an associated RF kernel.

The idea of kernel characterization of a random forest emerged as early as random forests themselves [14]. The most direct approach is to define a similarity from the frequency at which two samples end up in the same terminal node (leaf) of a tree. In particular, the limit of the probability two samples share a leaf in a

random forest

$$K(x_i, x_j) = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \sum_{t=1}^T I(x_i, x_j \in R_t(\theta_m)) \quad (4.5)$$

defines a positive definite kernel [14] (M - number of trees in the forest, T - number of leaves in an individual tree, θ_m - random vector of indices describing the training set for the m th tree, R_t - t th terminal rectangle (leaf) of the m th tree, $I(\cdot)$ - indicator function). The positive definiteness of the kernel is maintained if a cut at a random height in each tree (with its associated data point partitioning) is used in place of the partitioning induced by the tree leaves (the latter is just a special case with the cut set at zero height)[40]. Furthermore, kernels generated from extremely randomized trees (ERT) and their finite approximations are discussed in [53] and [14], while [127] suggests an updated kernel construction weighting scheme.

Our definition of an RF kernel extends (4.5) by incorporating two additional RF-derived statistics (Fig.4.15). First, if two data points have similar predicted labels across the trees of an RF, they should be considered alike even if they don't often fall in the same leaves. Furthermore, earlier node splits in a tree generally separate more distinct data point groups, while later splits tend to fine tune the decision boundary and often highlight much smaller differences between populations. Therefore, two data points that end up in one early- and one late-split leaf should be considered more distinct than data points in two late-split leaves, even if the latter do not have matching predicted labels.

Our 3-input RF kernel is computed in the following manner:

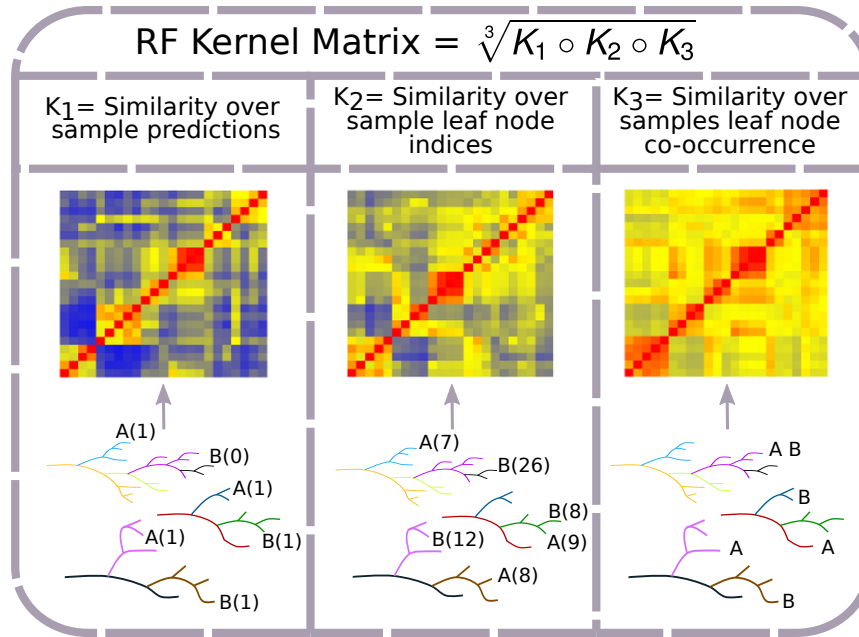


Figure 4.15: RF kernel construction. The RF Gram matrix is a geometric mean of the Hadamard product of three component similarity matrices, each capturing a different aspect of the RF model for a particular feature set: K_1 - similarity over RF tree label predictions (in the case of classification, probability of belonging to a class), K_2 - similarity over RF tree sample leaf indices, K_3 - proportion of times samples are assigned to the same RF tree leaf.

1. Calculate similarity over predictions across RF trees

$$K_1(x_i, x_j) = \exp\left(-\frac{\|p_i - p_j\|^2}{\sigma}\right) \quad (4.6)$$

where $p_i = (p_{i,1}, p_{i,2}, \dots, p_{i,M})$ is a vector of predictions for data point i from the M trees in the RF. p_i always has continuous entries - either the actual predictions in a regression setting, or the probabilities of class membership for classification problems. We set $\sigma = \max_{i,j} \|p_i - p_j\|^2$ so that distances are scaled to the $[0, 1]$ range. The distance is exponentiated to convert to similarity.

2. Calculate similarity over leaf node indices across RF trees

$$K_2(x_i, x_j) = \exp\left(-\frac{\|t_i - t_j\|^2}{\sigma}\right) \quad (4.7)$$

where $t_i = (t_{i,1}, t_{i,2}, \dots, t_{i,M})$ is a vector of the indices of data point i leaf nodes across the RF. Tree nodes are indexed starting from the base node and then sequentially across each depth level of the tree. We set σ as in K_1 .

3. Calculate similarity over probability of leaf node co-occurrence by using the finite approximation of Eqn.(4.5)

$$K_3(x_i, x_j) = \exp\left(\left(\frac{1}{M} \sum_{m=1}^M \sum_{t=1}^T I(x_i, x_j \in R_t(\theta_m))\right) - 1\right) \quad (4.8)$$

We add the exponential transformation to keep K_3 on a similar scale as K_1 and K_2 in order to avoid any one component having a disproportionate effect.

4. Calculate the composite kernel as the geometric mean of the element-wise product of K_1, K_2 and K_3

$$K(x_i, x_j) = \sqrt[3]{K_1(x_i, x_j) \circ K_2(x_i, x_j) \circ K_3(x_i, x_j)} \quad (4.9)$$

It is important to note that K as well as all component kernels K_1, K_2 and K_3 are positive definite. Indeed, K_1 and K_2 are Gaussian kernels over $\mathcal{P} \times \mathcal{P}, \mathcal{P} \subseteq \mathcal{I}^+$ and $\mathcal{T} \times \mathcal{T}, \mathcal{T} \subseteq \mathcal{I}^+$ respectively, which ensures their positive definiteness [131]. Furthermore, K_3 is the exponent of a positive-definite kernel, i.e. a positive-definite kernel as well [131]. Finally, K is the Hadamard product of three positive-definite kernels raised to a positive power - since both of these operations preserve positive-definiteness [131], K is a proper kernel itself.

4.5.5 Stacked Learning

Stacked generalization is a special case of ensemble learning in which the ensemble, or metalearner, uses the output of its components (base learners) to learn the final predictions [169][13]. What makes stacked learning unique is that the base learner predictions (level-one data) are generated in a cross-validated manner that excludes each sample from the training set that produces its predicted label. More specifically, if our training set (level-zero data) is $X = \{X_i : i = 1 \dots N, X_i \in \mathcal{R}^p\}$ with labels $Y = \{Y_i : i = 1, \dots, N, Y_i \in \mathcal{R}\}$ and we have a collection of base learners $\{\Delta_1, \dots, \Delta_S\}$ then stacked generalization proceeds as follows [111][89]:

1. Randomly split the level-zero data into K folds of roughly equal size - $v_1, \dots, v_k$ (K -fold cross validation).
2. For each of the Δ_s base learners, train k models (collectively denoted as

$\hat{\Delta}_{s,cv}$). For each model, use $X \setminus X_{v_k}$ to train on; use same model to generate predictions for X_{v_k} .

3. Concatenate the k sets of predictions for each base learner into a vector z of length N . The $N \times S$ matrix Z of such vectors for all S base learners is the level-one training set.
4. Train a metalearner Ψ on Z (with hyperparameter tuning, if necessary).
5. Train each Δ_s base learner on the full level-zero training set X to obtain $\hat{\Delta}_s$. The stacked model is $\{\hat{\Delta}_s : s = 1, \dots, S; \Psi\}$. To predict on a new data point X_{new} , we compute a $1 \times S$ vector $Z_{new} = \{\hat{\Delta}_s(X_{new}) : s = 1, \dots, S\}$ and use $\Psi(Z_{new})$ as the stacked model's prediction.

4.5.5.1 Super Learner

Stacked learning has no specification requirements on the metalearner Ψ . The disadvantage of such flexibility is the lack of theoretical results for the improved empirical performance of stacking. A Super Learner [159] is a type of stacked learner with restrictions on Ψ so that desirable theoretical properties can be proven. The main such condition is that the optimal metalearner Ψ^* is the minimizer of a bounded loss function. A Super Learner for which $\{\Delta_1, \dots, \Delta_S\}$ and Ψ have uniformly bounded loss functions exhibits the "oracle" property - Ψ^* is asymptotically guaranteed to perform as well as the optimal base learner Δ_s^* under the true data-generating distribution [158][156]. Furthermore, if we constrain the choice of Ψ to $\Psi = \sum_{i=1}^S \alpha_i \Delta_i, \forall \alpha_i \geq 0$, Ψ^* asymptotically converges to the performance of the optimal conical combination of $\{\Delta_1, \dots, \Delta_S\}$ [159][111]. This theoretical result is a main reason why in practice Ψ is often chosen to be some

form of regularized linear or logistic regression.

For example, if the aim is to predict a continuous variable (regression task), one can set $\Psi = \sum_{i=1}^S \alpha_i \hat{\Delta}_{i,cv}$ and solve the regression problem

$$\min_{\alpha} \sum_{i=1}^N (Y_i - \sum_{j=1}^S \alpha_j \hat{\Delta}_{j,cv}(X_i))^2 \quad (4.10)$$

possibly regularized or subject to a convex constraint on the α weights ($\forall \alpha_i \geq 0, \sum_{i=1}^S \alpha_i = 1$)[159] [111].

Similarly, for classification problems the squared error loss of (4.10) can be replaced with the logistic loss:

$$\min_{\alpha} \sum_{i=1}^N \log(1 + \exp(-Y_i * (\sum_{j=1}^S \alpha_j \hat{\Delta}_{j,cv}(X_i)))) \quad (4.11)$$

and all theoretical Super Learner results will still hold.

4.5.6 AKLIMATE as a Super Learner

AKLIMATE is an example of a Super Learner - the base learners $\{\Delta_1, \dots, \Delta_S\}$ are RFs trained on biologically relevant feature sets, and the metalearner Ψ is a regularized MKL, the latter being a linear regression in a new (unobservable) feature space. To the best of our knowledge, this is the first time a kernel-based stacked learning framework has been proposed.

Some of the most salient ways that AKLIMATE incorporates Super Learner characteristics are:

1. To maximize performance, the base learners should be as diverse as possible - either by using different algorithms, or by varying the parameters of a par-

ticular modeling approach [13][159]. AKLIMATE achieves diversity through the use of different feature subsets for each RF base learner.

2. The AKLIMATE equivalent to the level-one Z training set are level-one kernels. To compute them efficiently, AKLIMATE makes use of RF's inherent ability to generate OOB predictions (predictions computed from trees for which the data point in question was not part of the training set). Such OOB predictions serve the same purpose as the "regular" CV-based level-one training data and can successfully replace the latter. The level-one RF kernels (OOB RF kernels) are thus computed as follows:

- (a) For a RF with M trees, define $OOB(m)$ as the set of data points that are OOB in the m th tree. Furthermore, let

$$I_{OOB}(x_i, x_j) = (\delta_{1ij}, \dots, \delta_{mij}) \text{ where} \quad (4.12)$$

$$\delta_{mij} = \begin{cases} 1 & x_i, x_j \in OOB(m), \\ 0 & \text{otherwise} \end{cases}$$

i.e. I_{OOB} is the tree-level indexing function of when x_i and x_j are simultaneously OOB in a given RF.

- (b)

$$K_1(x_i, x_j) = \exp\left(-\frac{\|\langle p_i, I_{OOB}(x_i, x_j) \rangle - \langle p_j, I_{OOB}(x_i, x_j) \rangle\|^2}{\sigma}\right) \quad (4.13)$$

- (c)

$$K_2(x_i, x_j) = \exp\left(-\frac{\|\langle t_i, I_{OOB}(x_i, x_j) \rangle - \langle t_j, I_{OOB}(x_i, x_j) \rangle\|^2}{\sigma}\right) \quad (4.14)$$

(d)

$$K_3(x_i, x_j) = \exp\left(\left(\frac{1}{\sum I_{OOB}(x_i, x_j)} \sum_{m=1}^M \sum_{t=1}^T I(x_i, x_j \in R_t(\theta_m)) I_{OOB}(x_i, x_j)_m\right) - 1\right) \quad (4.15)$$

(e)

$$K(x_i, x_j) = \sqrt[3]{K_1(x_i, x_j) \circ K_2(x_i, x_j) \circ K_3(x_i, x_j)} \quad (4.16)$$

with the same notation and σ calculations as in (4.6)-(4.9).

3. AKLIMATE's MKL metalearner has two elastic-net hyperparameters (λ_1, λ_2) that require tuning. We do so by generating a random set of (λ_1, λ_2) pairs and ranking them based on a K -fold (default $K = 5$) cross-validation fit metric with OOB RF kernels as input. To improve generalization, we use a simplified version of the overfit correction procedure in [108] - instead of selecting the hyperparameters that produce the best CV fit, we choose the ones corresponding to the 90th percentile of the distribution of the CV fit metric.

4.5.7 AKLIMATE algorithm

Let our training set be $(X, Y) = \{(X_i, Y_i) : X_i = X_{i1} \cup \dots \cup X_{id}, i = 1 \dots N, d = 1 \dots D\}$ where we have N data points and D data types with feature memberships $\{C_i\}_{i=1}^D$ respectively. Furthermore, let there be S feature sets, each containing a list of features $P(s)$. We can now formally define AKLIMATE (Algorithm 1):

Algorithm 1: AKLIMATE

```
for  $s \in S$  do
     $\widehat{P}(s) = \cup_{d=1}^D (P(s) \cap C_d)$  ;           // list of available features
     $RF_s(\widehat{P}(s), Y)$  ;                       // Train respective base learner
     $\widehat{Y}^s = RF_s(\widehat{P}(s), Y)(X)$  ;           // Compute OOB predictions
     $\tau_s = \text{Fit}(Y, \widehat{Y}^s)$  ;           // fit statistic based on OOB predictions
end
Sort  $\{RF_s(\widehat{P}(s), Y)\}_{s=1}^D$  by  $\{\tau_s\}_{s=1}^S$ 
for  $n \in N$  do
     $BestRF_n = \{\{RF_k(\widehat{P}(s), Y)\}_{k=1}^K : \widehat{Y}_n^s = Y_n \ \& \ \tau_1 \geq \tau_2 \cdots \geq \tau_k \geq$   

     $\max(\{\tau_s\}_{s=1}^S \setminus \{\tau_1, \dots, \tau_k\})\}$  ; // pick top  $k$  (ranked by  $\tau_s$ ) RFs
    that predicted data point  $n$  correctly
end
 $BestRFs = \cup_{n=1}^N BestRF_n$ 
Sort  $BestRFs$  by  $\tau_s$ 
for  $r \in BestRFs$  do
     $K_r = \text{RF OOB Kernel}(r)$ 
end
 $K_{oob} = \{K_r\}_{r=1}^R$ 
 $(\lambda_1^*, \lambda_2^*) = \text{MKL CV}(K_{oob}, Y)$  ;           // Tune hyperparameters
for  $r \in BestRFs$  do
     $K_r = \text{RF Kernel}(r)$ 
end
 $K_{full} = \{K_r\}_{r=1}^R$ 
 $\Psi = \text{MKL}(K_{full}, \lambda_1^*, \lambda_2^*, Y)$  ;           // Train MKL metalearner
```

The *BestRF* selection step in algorithm 1 describes AKLIMATE’s behavior for a binary/categorical Y (a classification task). If Y is continuous (regression problem) the predictions and labels cannot be directly compared for equality. One approach is to take the mean squared error of the prediction-label differences and use that metric to re-rank RFs for each sample - in our experience this leads to the selection of suboptimal RFs due to overfitting. Instead, we propose a more robust scheme that shows better results in practice - we binarize the vector of predictions

$\{\hat{Y}_n^s\}_{s=1}^S$ for each data point into matching and non-matching predictions and then apply the standard classification case selection rule. The binarization of the prediction of a particular RF for data point n is done as follows:

$$\hat{Y}_{n_{bin}}^s = \begin{cases} 1 & \text{if } |\hat{Y}_n^s - Y_n| \leq \text{quantile}(\{|\hat{Y}_n^s - Y_n|\}_{s=1}^S, q), \\ 0 & \text{otherwise} \end{cases} \quad (4.17)$$

where q is a user-specified quantile of the empirical distribution of $\{|\hat{Y}_n^s - Y_n|\}_{s=1}^S$ (default $q = 0.05$). Our setup prioritizes RFs that perform near-optimally on individual data points and optimally when the training set is considered as a whole.

It should be emphasized that the *BestRF* selection step is the reason AKLIMATE can query significantly more kernelized gene sets than MPL. SpicyMKL and Spicer (see 4.5.10) require all kernels to be pre-computed as MKL optimization requires their simultaneous presence. When the number of gene sets is in the thousands, this can lead to very slow MKL optimization times, even for small data sets. AKLIMATE's *BestRF* step filters the full collection of feature sets (tens of thousands of entities) down to a subgroup of relevant sets that can be two or more orders of magnitude smaller in size. It is effectively a parallelizable univariate feature set relevance ranking. Since it is often the case that only a small proportion of the feature set collection is truly explanatory for a given prediction task, filtering out the non-relevant parts of the compendium does not impact negatively the multivariate analysis step (MKL) yet produces a drastic improvement in computational time.

Finally, the *BestRF* ranking step allows for the selection of optimal data types for each base RF - different (user-supplied) data type combinations can

be considered during learning and evaluated on their OOB performance. This is particularly useful if some of the queried feature sets are data-type specific. Furthermore, our RF filtering setup can also allow for individual hyperparameter tuning (e.g. minimum node size, *mtry*) for each component RF, although care should be taken to avoid overfitting..

4.5.8 Feature and feature set weights

Feature set weights w^{FS} , $\sum_i w_i^{FS} = 1$ are recovered directly from the optimal MKL metalearner Ψ^* (Eqn.(4.4)). Feature weights can then be calculated as $w_i^F = \sum_{k \in \widehat{P(\cdot)}_i} w_k^{FS} m(RF_k, i)$ where $\widehat{P(\cdot)}_i = \{\widehat{P(s)} : i \in \widehat{P(s)}, s = 1, \dots, S\}$ is the set of all feature sets s that have i as its member, and $m(RF_k, i)$ is an RF -specific feature score computed from the k th RF with feature i among its input features.

The simplest way to compute $m(RF_k, i)$ is by averaging the improvement in the splitting criterion over all nodes that used feature i as splitting variable. For the often-used Gini impurity measure, this involves computing the mean difference in impurity before and after each split, with larger mean impurity decreases indicative of more important variables [16]. While fast, this version of $m(RF_k, i)$ suffers from important shortcomings such as bias in favor of variables with more potential split points (e.g. continuous or categorical with a large number of categories), particularly when trees train on bootstrapped data [138]. Many alternatives for $m(RF_k, i)$ have been proposed [138][120][2]. For our work, we chose as default the permutation-based importance calculation described in the original RF paper [15] - the vector of measurements for feature i is randomly permuted and the permuted variable is used in the calculation of OOB predictions; the difference in error rate between the permuted and non-permuted OOB predictions is taken as

a measure of feature i 's importance. Permutation-based importance is intuitive, robust and generally performs as well as or just slightly worse than newer more complex $m(RF_k, i)$'s. Its biggest drawback is the higher computational cost. In cases where compute time is the main constraint, we recommend the actual impurity reduction (AIR) metric [106] - an extension of the pseudodata-augmented approach in [120]. It is similar in speed to pure Gini impurity importance, but retains the desirable properties of perturbation-based methods. While AIR can lead to a small prediction accuracy penalty due to the way it integrates variable importance computations in the RF trees construction, in our experience this effect has been negligible for classification tasks (for regression tasks we still recommend permutation-based importance).

4.5.9 AKLIMATE hyperparameter settings

We ran AKLIMATE with the same gene set collections across all prediction tasks (see 4.4.4). To increase robustness, we discarded gene sets that had fewer than 15 features across all data types considered - since different tasks use different number of data types, such thresholding causes the number of eligible gene sets to be task-specific.

In addition, we used the same AKLIMATE hyperparameters in all case studies. AKLIMATE component RFs used sampling without replacement, with each tree trained on a randomly selected 50% subsample of the training set. We chose this subsampling scheme both because it reduces computational time and because it performs as well as bootstrapping in predictive accuracy benchmarks [50]. In addition, the trees in each RF base model had minimum leaf size set to 1% of the size of the cohort (i.e. they were nearly fully grown trees). For the RF base

model ranking step (see 4.5.7) each forest contained 500 trees, while for the kernel construction step the forests grew to 2000 trees.

The number of features randomly selected to try at each node (*mtry*) was set to 25% of the size of the queried gene set. It is generally recommended to keep *mtry* low - e.g. $\sqrt{P}$ (P - total number of features) for classification or $\frac{P}{3}$ for regression problems [92]. Lower values of *mtry* tend to reduce correlation among predictions of individual trees at the expense of increased variance of error estimates. However, due to the averaging effect of ensembles of decorrelated models, they tend to reduce the error variance of the ensemble RF [94]. That effect is somewhat negated by an increase in the bias of RF error estimates, but generally speaking the more decorrelated the trees the better the RF performance [94]. AKLIMATE, however, is an ensemble of ensembles - variance reduction through base model decorrelation can also be achieved by selecting RFs that describe independent gene sets. As a consequence, we can prioritize bias reduction within RFs - we recommend *mtry* values in the 25-75% range, similar to or higher than what is normally recommended for RF regression tasks (33%)[92]. In our experience, AKLIMATE's default setting of 25% is computationally fast and causes negligible loss of performance.

Finally, we use two different importance metrics for the calculation of feature and feature set relevance - actual impurity reduction (AIR) [106] and permutation analysis [15] (see 4.5.8). The latter is robust and applicable across all prediction tasks, which is why we recommend it as a default setting. AIR, however, performs just as well in classification problems, with a noticeable reduction in computation time - for that reason we used it for our microsatellite instability and metabolic survival analyses.

4.5.10 Implementation

We use the R package *ranger* [171] for calculations involving AKLIMATE’s base RF learners, including permutation-/AIR-based variable importance. We chose *ranger* for its flexible handling of multiple splitting rules, variable importance approaches, and learning tasks; it is also one of the fastest RF algorithms currently available, particularly in problems where the number of features is much larger than the number of data points.

For our MKL learner we ported *SpicyMKL* [145] to R. We chose *SpicyMKL* because its proximal minimization formulation guarantees super-linear optimization convergence - the boost in speed makes it possible to handle 1,000’s of kernels. Furthermore, *SpicyMKL*’s elastic-net regularization allows maximum flexibility in terms of the number of kernels included in the optimal solution. Our R implementation of *SpicyMKL* called *Spicer* is available at <https://github.com/VladoUzunangelov/Spicer>.

Finally, an R implementation of AKLIMATE is available at <https://github.com/VladoUzunangelov/AKLIMATE>.

Chapter 5

Future Directions

Chapter 4 demonstrated AKLIMATE’s flexibility in dealing with various combinations of clinical and molecular data. In this chapter, I highlight AKLIMATE’s extensibility. I briefly describe three ways in which AKLIMATE can be augmented (one of which is already developed) as well as outline how AKLIMATE can be used to construct the pathway-pathway dependency map that my MPL shRNA knockdown analysis (Chapter 3) was building towards.

5.1 Multiclass AKLIMATE

AKLIMATE can be easily extended to handle multiclass classification problems. In fact, the only modification of its standard binary classification setup is the extension of the MKL metalearner to multiclass analysis. Multiclass MKL can be approached in several ways. One can implement a multiclass logit loss function as suggested in [151] and solve a version of the Dual Augmented Lagrangian problem in [145]. Alternatively, one can construct a wrapper around a binary classification MKL that trains a separate submodel for each pair of

label classes then aggregates the results of all such submodels. I chose the latter option - my implementation of multiclass MKL is available at <https://github.com/VladoUzunangelov/Spicer>.

5.1.1 Minimal Feature Predictors of TCGA Subtypes

Multiclass AKLIMATE can be applied to the development of minimal feature predictors for patient subtypes across TCGA projects. TCGA subtypes were discovered through the analysis of well-curated multi-modal data. However, at present it is difficult to use them clinically - to classify a new patient into an existing subtype a physician will need measurements for tens of thousands of genomic and transcriptomic features, which is currently both time consuming and costly. If accurate predictive models for TCGA subtypes existed, they could pinpoint the most discriminative features and greatly reduce data collection cost. To this end, the TCGA Tumor Molecular Pathology working group (TMP AWG, of which I am a member) aims to create predictive models for TCGA subtypes that maintain high accuracy but depend on a minimal number of features (preferably ~ 1000 , shared by TCGA subtypes across all 33 working groups). Ideally such models will be simple and easy to interrogate - for example, the gastric adenocarcinoma subtypes [107] can be accurately recapitulated with only 4 features (although 2 of them are composites aggregating information from whole data types, Fig. 5.1). The combined set of features across all minimal feature predictors can then be measured on a custom array, providing a cheap and fast method for a clinician to place a new patient in an existing molecular subtype.

AKLIMATE is well suited for the creation of such predictive models due to its ability to integrate multimodal data. I will now describe some results from

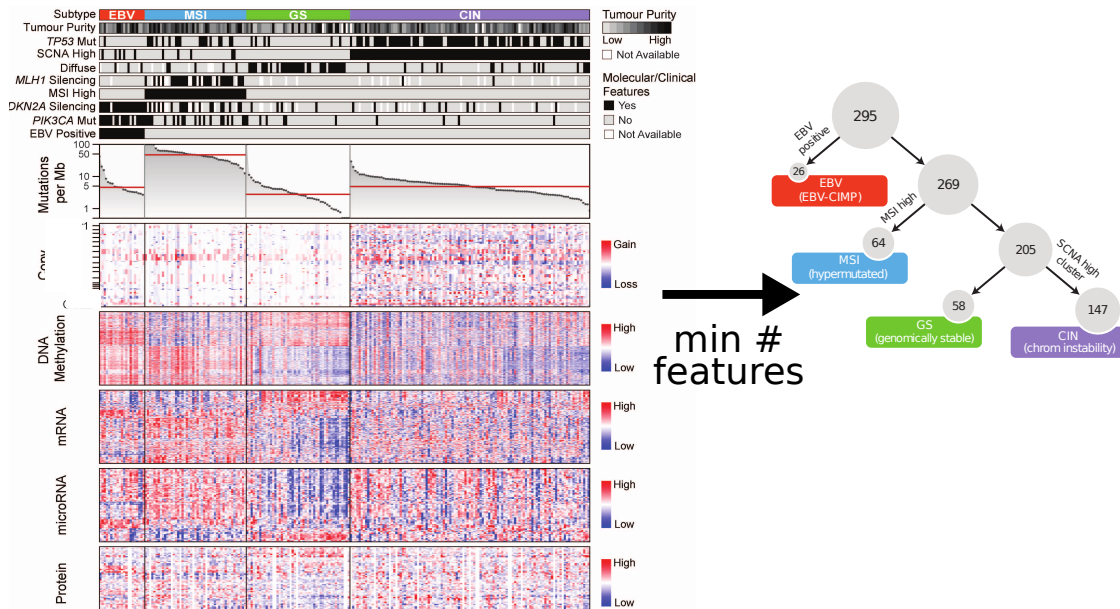


Figure 5.1: Minimal Feature predictors of TCGA subtypes. TCGA subtypes are defined on the basis of tens of thousands of features, sometimes spanning several different platforms. However, the subtypes can often be characterized by a handful of covariates that can predict each subtype with a high degree of accuracy. For example, the four subtypes of gastric adenocarcinoma identified in the TCGA marker paper [107] can be faithfully recapitulated by two clinical variables (EBV+, diffuse histological subtype) and two composite features (MSI, chromosome instability). Figure panels taken from [107].

the pilot TMP AWG analysis of the liver hepatocellular (LIHC) and cholangiocarcinoma (CHOL) TCGA cohorts. All data collection and clustering results were performed by other members of the working group while I contributed the AKLIMATE analysis.

Since the publication of the LIHC TCGA marker paper [168], the TCGA LIHC cohort has grown almost two-fold (from 196 to 379 samples). Led by the desire to create subtypes that are as comprehensive as possible, we repeated the analysis in [168] with the larger cohort now available. Furthermore, since cholangiocarcinoma (bile duct cancer) affects another type of liver cells and there is a known LIHC subtype that is molecularly similar to it, we decided to merge the LIHC and CHOL[49] cohorts for the purpose of re-running our clustering analysis. We settled on an iCluster [132] integrative clustering solution for 380 samples based on 4715 RNA, 4712 CN, 4705 methylation, 81 miRNA, and 691 lncRNA features (some samples were excluded because they were missing data in one or more of the component data types). A hierarchical clustering of the iCluster latent variables identified 7 subtypes - the associated dendrogram allowed us to construct a decision tree in which more distinct clusters split earlier and less distinct ones later (Fig. 5.2). We further characterized each cluster by a molecular or composite feature that was significantly enriched in that cluster and known to be biologically relevant. Finally, to make our prediction task easier, we eliminated some of the atypical cluster members by filtering out the bottom 10% by silhouette score [119]. All subsequent classification results describe that filtered clustering solution (334 samples).

Encouraged by the accuracy of the "most salient feature" decision tree approach in recapitulating gastric subtypes (Fig. 5.1), I attempted to predict

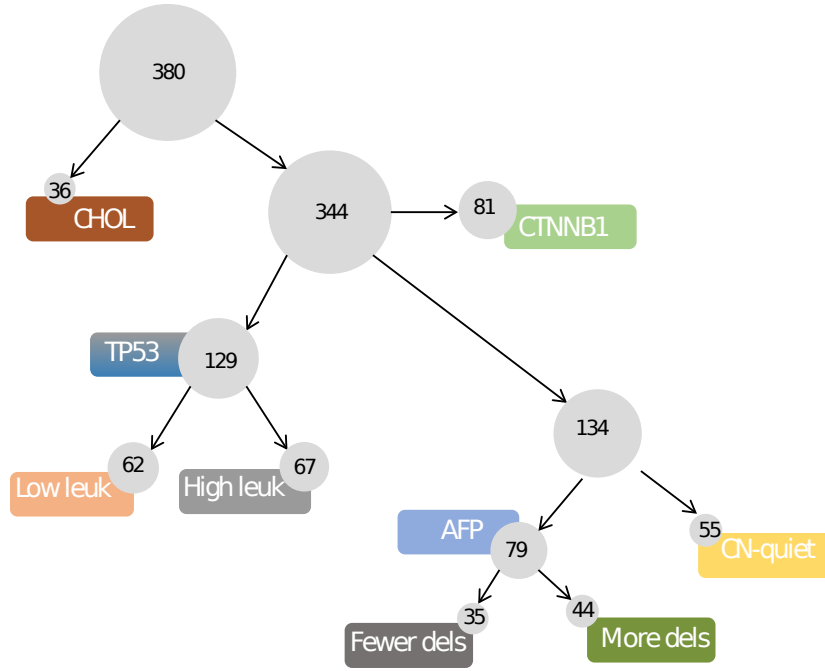


Figure 5.2: Subtype decision tree for the iCluster integrative clustering solution of the combined TCGA+CHOL cohort. The decision tree is derived from a hierarchical clustering dendrogram using iCluster latent variables as features. Numbers within each node state the size of the sample group available for that split. Earliest splitting nodes represent clusters that are most dissimilar from the rest, while later splits indicate less dissimilarity between groups. Each cluster is named after a clinical or molecular feature that it is highly enriched for. Figure created by Ronglai Shen and Esther Drill for the TCGA TMP AWG.

cluster membership through the use of 8 significantly enriched biomarker features that gave the LICH+CHOL clusters their names (CDKN2A silencing; IDH, TP53, BAP1 and CTNNB1 mutations; alpha feto-protein levels, leukocyte fraction, altered genome fraction). Unfortunately, models limited to the use of these 8 features were insufficiently accurate (mean overall balanced accuracy (MBA) 0.717 ± 0.01 , Fig. 5.3A). The performance was particularly bad for some of the later-splitting clusters (AFP-fewer deletions - MBA 0.517 ± 0.019 ; TP53- low leukocyte fraction - MBA 0.658 ± 0.021), indicating a need for the inclusion of additional molecular features.

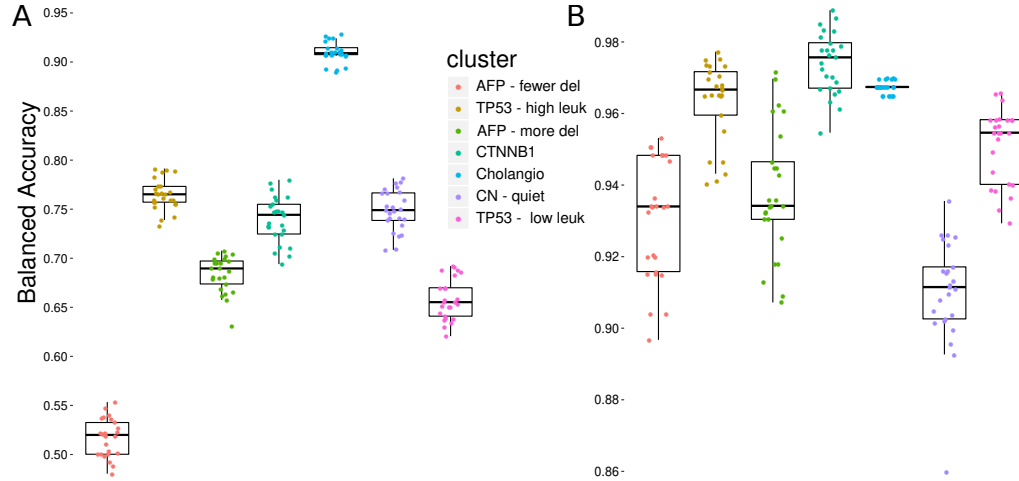


Figure 5.3: Performance of predictive models using A) 8 significantly enriched biomarker features and B) full AKLIMATE models. Models in A) were Random Forests grown to maximum tree depth, with 5000 trees and default mtry (rounded $\sqrt{8}$). Models in B) were AKLIMATE runs with the same hyperparameters, pathway collections and filtering steps described in Chapter 4. Barplots describe per-subtype balanced accuracies over 25 repetitions of 5 fold cross-validation (same folds in A) and B)). Balanced accuracy computed by taking the average of the sensitivity and specificity for each subtype.

On the other hand, an AKLIMATE model trained on $>65,000$ molecular features (Table 5.1) achieves an MBA of 0.948 ± 0.006 , with no cluster predicted at

less than 90% MBA (Fig. 5.3B). As expected, models trained on more extensive sets of molecular features lead to drastically improved predictive performance. It is unclear, however, whether the full model accuracy can be achieved with > 8 but $\ll 65,000$ features. To investigate the relationship between the number of input features and model accuracy, I devised the following experiment:

1. For each full AKLIMATE model in Fig. 5.3B, I created a feature importance ranking as described in Section 4.5.8.
2. A Random Forest model was trained with the top K most important AKLIMATE features, $K = 5, 10, 20, 50, 100, 200, 500, 1000, 1500$.
3. Each reduced RF model was evaluated by measuring its balanced accuracy on the test fold not seen by the full AKLIMATE model.

Table 5.1: Number of features used in the full AKLIMATE model, by data type.

Data Type	Number Features
mRNA	12,393
methylation	34,139
CN gene level	11,036
CN regions	4,114
mutation	3,481
lncRNA	691

The results of that analysis are summarized in Fig. 5.4. The 5-feature reduced model with AKLIMATE-recommended features performs marginally better than the 8-biomarker RF models from Fig.5.3A (MBA 0.724 ± 0.017 vs 0.717 ± 0.01). The difference in performance becomes significant if we move to a 10-feature reduced model (MBA 0.794 ± 0.012) or an extrapolated 8-feature reduced model (MBA ~ 0.765) suggesting that the AKLIMATE-recommended features are more

informative than the 8 biomarkers even if the same number of features are considered. This improvement is even more striking if we take into account that 2 of the 8 biomarkers (leukocyte fraction and altered genome fraction) are composite entities that aggregate information from several molecular features, making the 20-feature or even the 50-feature reduced model a fairer direct comparison. A possible explanation for the better accuracy of models with AKLIMATE-recommended features might be that the 8 biomarkers were selected by univariate significance testing, whereas the AKLIMATE features are prioritized in multivariate analysis. In addition, AKLIMATE-recommended features benefit from the strong informative prior provided by the gene sets AKLIMATE was trained with.

The slope of the graph of reduced model accuracies in Fig. 5.4 is clearly not constant - the addition of features leads to much larger gains when the number of input features is small, with improvements flattening out after ~ 500 features. This is quite encouraging - a small increase in the number of molecular features used can produce model accuracy comparable to the one of the full feature set. In fact, with just 50 features we can achieve MBA of 0.885 ± 0.01 . What is more, as these features are derived from the full AKLIMATE model, we can still interrogate the latter for a more complete view of the biological processes they represent.

To understand that interplay, let's examine Fig. 5.5A. It presents the top 50 most informative features and the top 20 most informative feature sets in the full AKLIMATE model (these are the same features used in the 50-feature reduced model). Many of the feature sets are directly related to liver function, the 8 key biomarkers, or LIHC initiation and progression. For example, the Metabolism pathway describes one of the main liver functions (metabolism regulation), as does the Leptin pathway (hunger regulation through adipose tissue signaling).

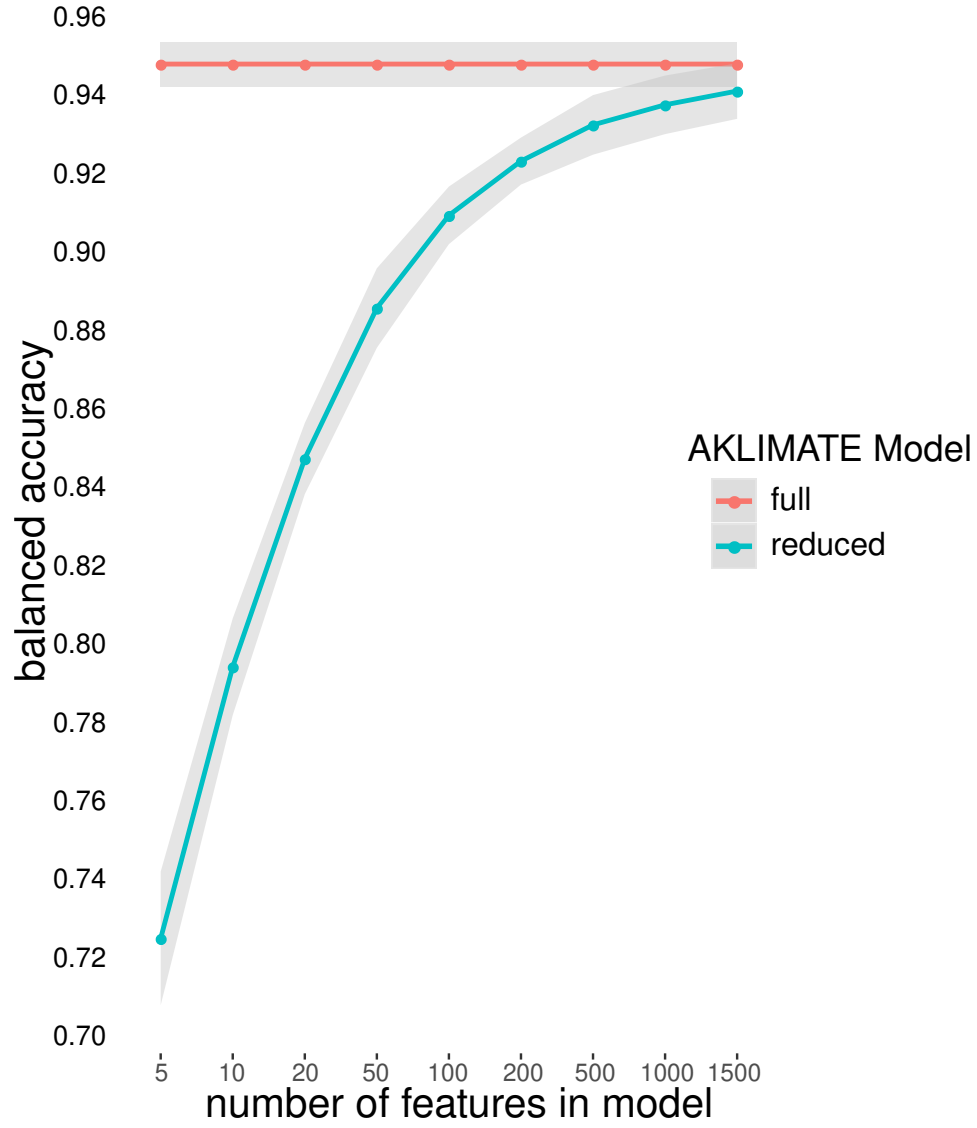


Figure 5.4: Comparison of balanced accuracy for full and reduced AKLIMATE models. The full models are trained on 25 repeats of 5 fold cross-validation splits, as described in Fig. 5.3B. The reduced models are RFs trained on the top K most predictive features from the full AKLIMATE model, for different values of K , using the same folds as the full AKLIMATE model. Each reduced model is evaluated on the data fold that the full AKLIMATE model did not see during training. The grey interval around each line denotes one standard deviation from the mean (each mean is denoted by a dot).

CTNNB1, a gene that is often mutated in LIHC[168] and one of the 8 original biomarkers, is also a key component of the WNT signaling pathway. In addition, WNT/CTNNB1 signaling is strongly involved in EGFR1 regulation in LIHC [146]. Furthermore, the TGF β and Inflammation pathways are directly related to inflammation response, as is leukocyte fraction (another one of the 8 original biomarkers).

An additional benefit of the full AKLIMATE model is that we can easily identify the core set of features shared among the most informative pathways. This core set is the nucleus of the dysregulated meta-pathway that AKLIMATE learns from its collection of available gene sets. It can be discerned from the pathway membership matrix in Fig. 5.5A, or more easily from the network representation of the same in Fig. 5.5B. For our LIHC+CHOL subtypes it appears that a key differentiating biological process revolves around MAPK (MAPK1,MAPK8) and PI-3 kinase (PIK3CA,PIK3R1) signaling. PIK3CA is a previously known driver of LIHC progression [168]. Less is known about PIK3R1’s involvement in liver cancer, although recent studies suggest it plays a key role in increasing proliferation and downregulating apoptosis [1].

Finally, we can also use the full AKLIMATE model to suggest features that contribute to the relevance of the top feature sets but have a slightly lower individual importance within the model. Such features can be useful substitutes for their more informative counterparts, particularly if they are highly relevant to other cancer subtypes - using them instead could help us keep the overall number of features down at the expense of a small decrease in accuracy for some prediction tasks. Figures 5.6 & 5.7 present expanded views into two of the key LIHC dysregulated pathways.

Figure 5.5: A) AKLIMATE results highlighting the top 20 most informative feature sets and top 50 most informative features for predicting iCluster LIHC+CHOL subtypes. Feature and feature set weights are averaged over 25 repeats of 5 fold cross-validation. Figure organized as Fig. 4.1. B) Network view of the top 10 feature sets and the top 50 features. Feature sets encoded as squares, features as circles, with different colors for different data types. The size of each node is proportional to the normalized weight assigned by AKLIMATE to the respective feature/feature set. This is a bipartite graph, with features linked to feature sets they are a member of. To improve interpretability, each feature set has a uniquely colored cloud that maps out its member features.

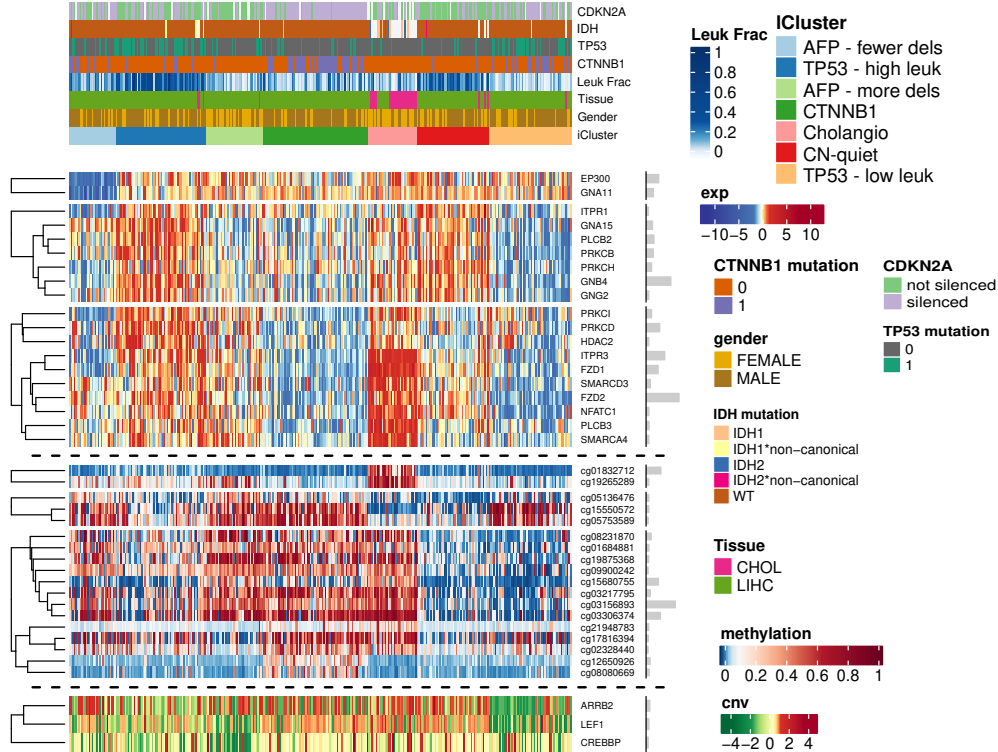


Figure 5.6: Extended set of informative features within the WNT signaling pathway. Features shown are members of the pathway that also score as one of the top 500 most relevant features overall. Figure organized as Fig. 4.1.

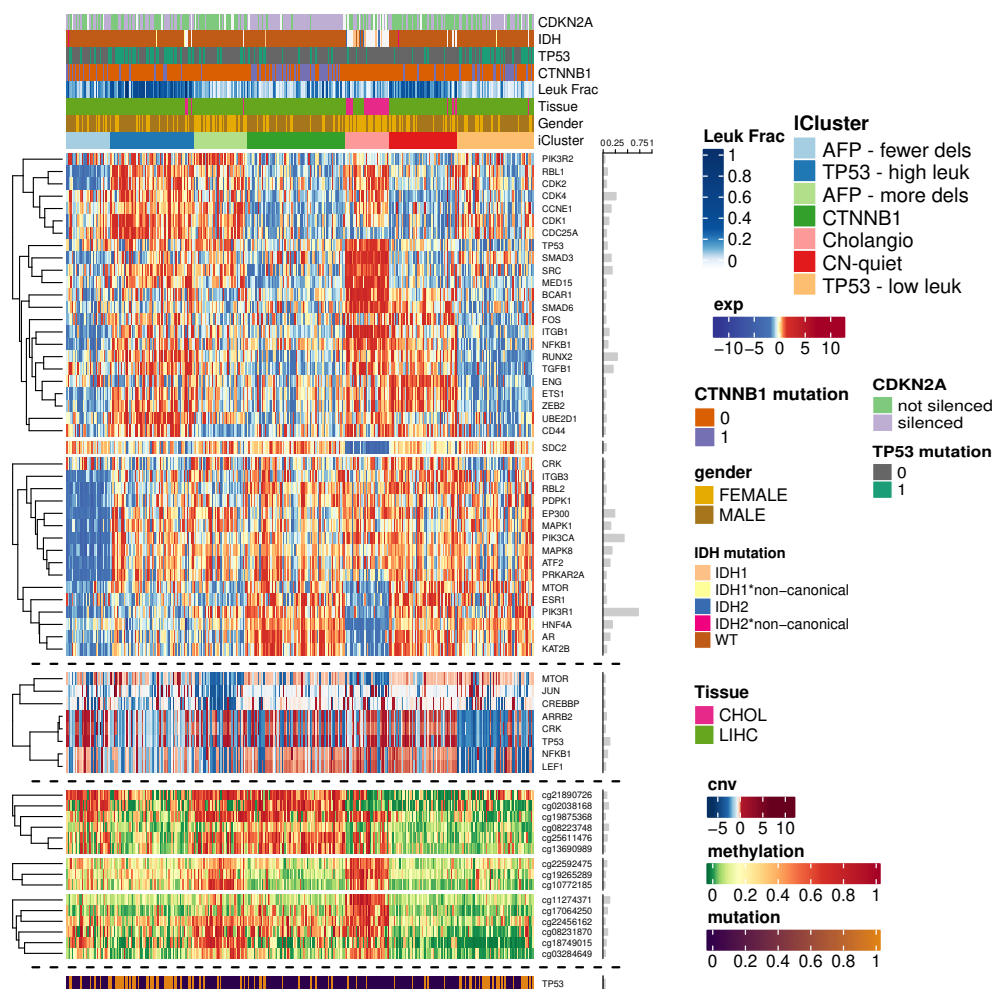


Figure 5.7: Extended set of informative features within the $TGF\beta$ signaling pathway. Features shown are members of the pathway that also score as one of the top 500 most relevant features overall. Figure organized as Fig. 4.1.

In conclusion, AKLIMATE can be of great service both as a multiclass classifier and as a feature prioritizer for minimum feature predictors.

5.2 Non-linear Stacked Learners

AKLIMATE uses a linear MKL as its stacked learning step. However, this choice is grounded in the desirable theoretical properties of linear combinations of kernels/base learners rather than any performance considerations. In fact, many competition-winning stacked learning approaches use a non-linear metalearner. Non-linear MKL solvers [31][56] or even multi-layer kernel machines (kernel learning in a deep learning framework) [25] are all very promising avenues to improve AKLIMATE performance.

Another potentially important upgrade would be the introduction of a joint learning model in place of the two separate learning steps for base and stacked models. In fact, if the random forest base learners and the deep learning kernel meta-learner can be jointly optimized in a manner similar to [83], it will likely lead to the largest gains in speed and accuracy.

5.3 Logical Dependency Features

While AKLIMATE utilizes prior knowledge in the form of feature sets, it does so in a "bag of features" way that does not incorporate known feature-feature dependencies - e.g. the feedback loop of MDM2-TP53 expression downregulation and TP53-MDM2 expression upregulation. Such very informative links, if explicitly modeled, will likely be the most important AKLIMATE improvement. One possible way to encode them is through feature engineering - we create new fea-

tures that encapsulate known logical dependencies. In the TP53-MDM2 case this can be achieved by adding two binary variables - one that is set to TRUE every time high MDM2 protein levels occur simultaneously with low TP53 expression and FALSE otherwise, and another one with similar settings for the TP53 protein - MDM2 transcripts relationship. Such an approach can capture a very diverse category of dependency structures as it is not limited to pairwise or same-type feature relationships.

5.4 Pathway-Pathway Dependency Map

Finally, AKLIMATE can be instrumental in constructing the pathway-pathway dependency map that proved elusive in Chapter 3. The main problem we encountered was the pervasiveness of significant feature set overlaps. Such overlaps confound attempts to compute KD-predictor pathway associations because the importance of a biologically significant process is partitioned among several gene sets that contain the key driver features of that process. Since AKLIMATE models compute both feature and feature set weights, we can use this dual representation to construct clusters of feature sets (arch-pathways) whose members share driver features. The introduction of this additional hierarchical level (the arch-pathway) effectively mitigates the feature set redundancy issue. As a consequence, the dependency map evolves from a bipartite graph with two types of nodes (KD and predictor pathways) to a quadripartite one with four node types (KD, predictor and arch-pathways, individual features), making it possible to capture significantly more sophisticated associations.

Bibliography

- [1] X. Ai et al. Overexpression of PIK3r1 promotes hepatocellular carcinoma progression. *Biological Research*, 51, Nov. 2018. ISSN 0716-9760.
- [2] A. Altmann, L. Tološi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10): 1340–1347, May 2010. ISSN 1367-4803.
- [3] M. J. Alvarez et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 48(8):838–847, Aug. 2016. ISSN 1061-4036.
- [4] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337, May 1950. ISSN 00029947.
- [5] M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000. ISSN 1061-4036.
- [6] A. Aytes et al. Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy. *Cancer Cell*, 25(5):638–651, May 2014. ISSN 1535-6108.
- [7] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Twenty-first international conference on Machine learning - ICML '04*, page 6, Banff, Alberta, Canada, 2004. ACM Press.
- [8] M. H. Bailey et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2):371–385.e18, Apr. 2018. ISSN 0092-8674, 1097-4172.
- [9] J. Barretina et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, Mar. 2012. ISSN 0028-0836.

- [10] T. Barrett et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Research*, 39(Database issue):D1005–D1010, Jan. 2011. ISSN 0305-1048.
- [11] T. Beer et al. Enzalutamide in Metastatic Prostate Cancer before Chemotherapy. *The New England journal of medicine*, 371(5):424–433, July 2014. ISSN 0028-4793.
- [12] E. Bilal et al. Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. *PLoS Computational Biology*, 9(5):e1003047, May 2013. ISSN 1553-7358.
- [13] L. Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, July 1996. ISSN 0885-6125, 1573-0565.
- [14] L. Breiman. Some infinity theory for predictor ensembles. Technical Report 577, UC Berkeley, 2000.
- [15] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. ISSN 1573-0565.
- [16] L. Breiman, J. H. Friedman, Olshen, R.A., and Stone, C.J. *Classification and Regression Trees*. Routledge, 1984. ISBN 978-1-351-46049-1.
- [17] I. H. G. Bronkhorst et al. Different Subsets of Tumor-Infiltrating Lymphocytes Correlate with Macrophage Influx and Monosomy 3 in Uveal Melanoma. *Investigative Ophthalmology & Visual Science*, 53(9):5370–5378, Aug. 2012. ISSN 1552-5783.
- [18] H. E. Bryant et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, 434(7035):913, Apr. 2005. ISSN 1476-4687.
- [19] C. Cai et al. Androgen receptor gene expression in prostate cancer is directly suppressed by the androgen receptor through recruitment of lysine-specific demethylase 1. *Cancer Cell*, 20(4):457–471, Oct. 2011. ISSN 1878-3686.
- [20] S. L. Carter et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, May 2012. ISSN 1546-1696.
- [21] E. G. Cerami et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(Database issue):D685–D690, Jan. 2011. ISSN 0305-1048.

- [22] C. Chattopadhyay et al. Uveal melanoma: From diagnosis to treatment and the science in between. *Cancer*, 122(15):2299–2312, 2016. ISSN 1097-0142.
- [23] W.-Y. Cheng, T.-H. O. Yang, and D. Anastassiou. Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. *Science Translational Medicine*, 5(181):181ra50–181ra50, Apr. 2013. ISSN 1946-6234, 1946-6242.
- [24] H.-S. Chiu et al. Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Reports*, 23(1):297–312.e12, Apr. 2018. ISSN 2211-1247.
- [25] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.
- [26] G. Ciriello et al. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, Oct. 2013. ISSN 1061-4036.
- [27] W. S. Cleveland, E. Grosse, and W. M. Shyu. *Local regression models*. JM Chambers, T. Hastie, eds. *Statistical Models in S*. Chapman and Hall/CRC, New York, 1992.
- [28] T. Colombo, L. Farina, G. Macino, and P. Paci. PVT1: a rising star among oncogenic long noncoding RNAs. *BioMed Research International*, 2015: 304208, 2015. ISSN 2314-6141.
- [29] G. O. Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, Jan. 2019. ISSN 0305-1048.
- [30] T. U. Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, Jan. 2015. ISSN 0305-1048, 1362-4962.
- [31] C. Cortes, M. Mohri, and A. Rostamizadeh. Learning Non-Linear Combinations of Kernels. *Journal of Machine Learning Research*, page 9, 2012.
- [32] J. C. Costello et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, Dec. 2014. ISSN 1546-1696.
- [33] S. E. Coupland, S. L. Lake, M. Zeschnigk, and B. E. Damato. Molecular pathology of uveal melanoma. *Eye*, 27(2):230–242, Feb. 2013. ISSN 0950-222X.

- [34] G. S. Cowley et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data*, 1:140035, Sept. 2014. ISSN 2052-4463.
- [35] C. J. Creighton et al. Development of resistance to targeted therapies transforms the clinically-associated molecular profile subtype of breast tumor xenografts. *Cancer research*, 68(18):7493–7501, Sept. 2008. ISSN 0008-5472.
- [36] C. J. Creighton et al. Molecular profiles of progesterone receptor loss in human breast tumors. *Breast cancer research and treatment*, 114(2):287–299, Mar. 2009. ISSN 0167-6806.
- [37] A. C. Culhane et al. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, 40(Database issue):D1060–D1066, Jan. 2012. ISSN 0305-1048.
- [38] C. Curtis et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, June 2012. ISSN 1476-4687.
- [39] B. Damato, J. A. Dopierala, and S. E. Coupland. Genotypic Profiling of 452 Choroidal Melanomas with Multiplex Ligation-Dependent Probe Amplification. *Clinical Cancer Research*, 16(24):6083–6092, Dec. 2010. ISSN 1078-0432, 1557-3265.
- [40] A. Davies and Z. Ghahramani. The Random Forest Kernel and other kernels for big data from random partitions. *arXiv:1402.4293 [cs, stat]*, Feb. 2014. arXiv: 1402.4293.
- [41] J. Dey et al. Voruciclib, a clinical stage oral CDK9 inhibitor, represses MCL-1 and sensitizes high-risk Diffuse Large B-cell Lymphoma to BCL2 inhibition. *Scientific Reports*, 7(1):18007, Dec. 2017. ISSN 2045-2322.
- [42] L. Ding et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, 173(2):305–320.e10, Apr. 2018. ISSN 0092-8674.
- [43] Y. Drew. The development of PARP inhibitors in ovarian cancer: from bench to bedside. *British Journal of Cancer*, 113 Suppl 1:S3–9, Dec. 2015. ISSN 1532-1827.
- [44] E. Efstathiou et al. Molecular Characterization of Enzalutamide-treated Bone Metastatic Castration-resistant Prostate Cancer. *European urology*, 67(1):53–60, Jan. 2015. ISSN 0302-2838.

- [45] P. Elyse S. Blum. Clinical Management of Uveal and Conjunctival Melanoma, Jan. 2016.
- [46] A. Fabregat et al. The Reactome pathway Knowledgebase. *Nucleic Acids Research*, 44(D1):D481–487, Jan. 2016. ISSN 1362-4962.
- [47] J. J. Faith et al. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol*, 5(1):e8, Jan. 2007.
- [48] H. Farmer et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, 434(7035):917, Apr. 2005. ISSN 1476-4687.
- [49] F. Farshidfar et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell reports*, 18(11):2780–2794, Mar. 2017. ISSN 2211-1247.
- [50] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, Feb. 2002. ISSN 0167-9473.
- [51] J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, Feb. 2010. ISSN 1548-7660.
- [52] L. Garcia-Alonso, M. M. Ibrahim, D. Turei, and J. Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *bioRxiv*, June 2018.
- [53] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, Apr. 2006. ISSN 0885-6125, 1573-0565.
- [54] A. E. Giuliano et al. Breast Cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, 67(4):290–303, 2017. ISSN 1542-4863.
- [55] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic acids research*, 36(Database issue):D154–158, Jan. 2008. ISSN 1362-4962.
- [56] Y. Gu et al. Nonlinear Multiple Kernel Learning With Multiple-Structure-Element Extended Morphological Profiles for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3235–3247, June 2016. ISSN 0196-2892.

- [57] M. Gönen. Integrating gene set analysis and nonlinear predictive modeling of disease phenotypes using a Bayesian multitask formulation. *BMC Bioinformatics*, 17(16):0, Dec. 2016. ISSN 1471-2105.
- [58] M. Gönen and E. Alpaydm. Multiple Kernel Learning Algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011. ISSN 1532-4435.
- [59] M. Gönen et al. A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of a Functional Screen of Cancer Cell Lines. *Cell Systems*, 5(5):485–497.e3, Nov. 2017. ISSN 2405-4712.
- [60] J. W. Harbour. A prognostic test to predict the risk of metastasis in uveal melanoma based on a 15-gene expression profile. *Methods in Molecular Biology (Clifton, N.J.)*, 1102:427–440, 2014. ISSN 1940-6029.
- [61] J. W. Harbour et al. Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science (New York, N.Y.)*, 330(6009):1410–1413, Dec. 2010. ISSN 1095-9203.
- [62] J. L. Haybittle et al. A prognostic index in primary breast cancer. *British Journal of Cancer*, 45(3):361–366, Mar. 1982. ISSN 0007-0920.
- [63] H. V. Heemers et al. Identification of a Clinically Relevant Androgen-Dependent Gene Signature in Prostate Cancer. *Cancer Research*, 71(5):1978–1988, Mar. 2011. ISSN 0008-5472, 1538-7445.
- [64] M. Hitchins et al. Dominantly Inherited Constitutional Epigenetic Silencing of MLH1 in a Cancer-Affected Family Is Linked to a Single Nucleotide Variant within the 5'UTR. *Cancer Cell*, 20(2):200–213, Aug. 2011. ISSN 1535-6108.
- [65] K. A. Hoadley et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944, Aug. 2014. ISSN 0092-8674, 1097-4172.
- [66] K. A. Hoadley et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6, Apr. 2018. ISSN 0092-8674, 1097-4172.
- [67] G. A. Hobbs, C. J. Der, and K. L. Rossman. RAS isoforms and mutations in cancer at a glance. *Journal of Cell Science*, 129(7):1287–1292, Apr. 2016. ISSN 0021-9533.

- [68] P. V. Hornbeck et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(D1):D512–D520, Jan. 2015. ISSN 0305-1048, 1362-4962.
- [69] S. Huang, K. Chaudhary, and L. X. Garmire. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, 8, 2017. ISSN 1664-8021.
- [70] N. Hunter and R. H. Borts. Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. *Genes & Development*, 11(12):1573–1582, June 1997. ISSN 0890-9369, 1549-5477.
- [71] I. H. Ismail et al. Germline mutations in BAP1 impair its function in DNA double-strand break repair. *Cancer Research*, 74(16):4282–4294, Aug. 2014. ISSN 1538-7445.
- [72] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press. ISBN 978-1-60558-516-1.
- [73] T. Jewison et al. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Research*, 42(Database issue):D478–484, Jan. 2014. ISSN 1362-4962.
- [74] R. N. Jorissen et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 14(24):8061–8069, Dec. 2008. ISSN 1078-0432.
- [75] H. Kalirai et al. Lack of BAP1 protein expression in uveal melanoma is associated with increased metastatic risk and has utility in routine prognostic testing. *British Journal of Cancer*, 111(7):1373–1380, Sept. 2014. ISSN 0007-0920.
- [76] M. Kanehisa et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, Jan. 2012. ISSN 0305-1048.
- [77] D. Kazmin et al. Linking ligand-induced alterations in androgen receptor structure to differential gene expression: a first step in the rational design of selective androgen receptor modulators. *Molecular Endocrinology (Baltimore, Md.)*, 20(6):1201–1217, June 2006. ISSN 0888-8809.

- [78] E. Khurana, Y. Fu, J. Chen, and M. Gerstein. Interpretation of Genomic Variants Using a Unified Biological Network Approach. *PLoS Computational Biology*, 9(3), Mar. 2013. ISSN 1553-734X.
- [79] M. Kim, N. Rai, V. Zorraquino, and I. Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nature Communications*, 7, Oct. 2016. ISSN 2041-1723.
- [80] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, Jan. 1971. ISSN 0022-247X.
- [81] M. Kloft, U. Brefeld, S. Sonnenburg, and P. Laskov. Efficient and Accurate p-Norm Multiple Kernel Learning. *Advances in Neural Information Processing Systems*, page 9, 2009.
- [82] M. Kloft, U. Rückert, and P. L. Bartlett. A Unifying View of Multiple Kernel Learning. In D. Hutchison et al., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6322, pages 66–81. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-15882-7 978-3-642-15883-4.
- [83] P. Kotschieder, M. Fiterau, A. Criminisi, and S. R. Buló. Deep Neural Decision Forests. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1467–1475, Santiago, Chile, Dec. 2015. IEEE. ISBN 978-1-4673-8391-2.
- [84] A. E. Koopmans et al. Clinical significance of immunohistochemistry for detection of BAP1 mutations in uveal melanoma. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc.*, 27(10):1321–1330, Oct. 2014. ISSN 1530-0285.
- [85] T. R. Kress, A. Sabò, and B. Amati. MYC: connecting selective transcriptional control to global RNA production. *Nature Reviews Cancer*, 15(10): 593–607, Oct. 2015. ISSN 1474-1768.
- [86] A. Lachmann et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26(19):2438–2444, Oct. 2010. ISSN 1367-4803.
- [87] N. Lallous et al. Functional analysis of androgen receptor mutations that confer anti-androgen resistance identified in circulating cell-free DNA from prostate cancer patients. *Genome Biology*, 17:10, Jan. 2016. ISSN 1474-760X.

- [88] G. R. G. Lanckriet et al. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, Nov. 2004. ISSN 1367-4803, 1460-2059.
- [89] E. LeDell. *Scalable Ensemble Learning and Computationally Efficient Variance Estimation*. PhD thesis, University of California, Berkeley, 2015.
- [90] B. X. Li et al. Systemic Inhibition of CREB is Well-tolerated in vivo. *Scientific Reports*, 6, Oct. 2016. ISSN 2045-2322.
- [91] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 2008. ISSN 1367-4803, 1460-2059.
- [92] A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2:5, 2002.
- [93] A. Liberzon et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011. ISSN 1367-4803.
- [94] G. Louppe. Understanding Random Forests: From Theory to Practice. *arXiv:1407.7502 [stat]*, July 2014. arXiv: 1407.7502.
- [95] S. Mallik and Z. Zhao. Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data. *Briefings in Bioinformatics*, 2019.
- [96] G. Manning. The Protein Kinase Complement of the Human Genome. *Science*, 298(5600):1912–1934, Dec. 2002. ISSN 00368075, 10959203.
- [97] D. Marbach et al. Wisdom of crowds for robust gene network inference. *Nat Meth*, 9(8):796–804, Aug. 2012. ISSN 1548-7091.
- [98] R. Marcotte et al. Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells. *Cancer Discovery*, 2(2):172–189, Feb. 2012. ISSN 2159-8274, 2159-8290.
- [99] R. Marcotte et al. Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell*, 164(1-2):293–309, Jan. 2016. ISSN 1097-4172.
- [100] A. A. Margolin et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006. ISSN 14712105. arXiv: q-bio/0410037.

- [101] A. A. Margolin et al. Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer. *Science translational medicine*, 5(181):181re1, Apr. 2013. ISSN 1946-6234.
- [102] C. McCarthy et al. Insights into genetic alterations of liver metastases from uveal melanoma. *Pigment Cell & Melanoma Research*, 29(1):60–67, 2016. ISSN 1755-148X.
- [103] P. Mendiratta et al. Genomic Strategy for Targeting Therapy in Castration-Resistant Prostate Cancer. *Journal of Clinical Oncology*, Mar. 2009.
- [104] C. H. Mermel et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011. ISSN 1465-6906.
- [105] Q. Mo et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, Mar. 2013. ISSN 0027-8424, 1091-6490.
- [106] S. Nembrini, I. R. König, M. N. Wright, and A. Valencia. The revival of the Gini importance? *Bioinformatics*, 2018.
- [107] T. C. G. A. R. Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202, Sept. 2014. ISSN 1476-4687.
- [108] A. Y. Ng. Preventing "Overfitting" of Cross-Validation Data. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 245–253, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-486-5.
- [109] A. Patnaik et al. A First-in-Human Phase I Study of the Oral p38 MAPK Inhibitor, Ralimetinib (LY2228820 Dimesylate), in Patients with Advanced Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 22(5):1095–1102, Mar. 2016. ISSN 1078-0432.
- [110] R. Petryszak et al. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*, 42(D1):D926–D932, Jan. 2014. ISSN 0305-1048, 1362-4962.
- [111] E. C. Polley. Super Learner In Prediction. Technical Report 266, UC Berkeley, 2010.

- [112] H. Poon, C. Quirk, C. DeZiel, and D. Heckerman. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*, page btu383, June 2014. ISSN 1367-4803, 1460-2059.
- [113] D. Pratt et al. NDEx, the Network Data Exchange. *Cell Systems*, 1(4): 302–305, Oct. 2015. ISSN 2405-4712.
- [114] D. Pratt et al. NDEx 2.0: A Clearinghouse for Research on Cancer Pathways. *Cancer Research*, 77(21):e58–e61, Nov. 2017. ISSN 0008-5472, 1538-7445.
- [115] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9(Nov):2491–2521, 2008. ISSN 1533-7928.
- [116] N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20): 10546–10562, Nov. 2018. ISSN 1362-4962.
- [117] A. G. Robertson et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell*, 32(2):204–220.e15, Aug. 2017. ISSN 1535-6108, 1878-3686.
- [118] D. Robinson et al. Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell*, 161(5):1215–1228, May 2015. ISSN 00928674.
- [119] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, Nov. 1987. ISSN 0377-0427.
- [120] M. Sandri and P. Zuccolotto. A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628, Sept. 2008. ISSN 1061-8600, 1537-2715.
- [121] C. F. Schaefer et al. PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(Database issue):D674–679, Jan. 2009. ISSN 1362-4962.
- [122] H. I. Scher et al. Antitumour activity of MDV3100 in castration-resistant prostate cancer: a phase 1-2 study. *Lancet (London, England)*, 375(9724): 1437–1446, Apr. 2010. ISSN 1474-547X.
- [123] H. I. Scher et al. Increased Survival with Enzalutamide in Prostate Cancer after Chemotherapy. *New England Journal of Medicine*, 367(13):1187–1197, Sept. 2012. ISSN 0028-4793, 1533-4406.

- [124] M. R. Schlabach et al. Cancer Proliferation Gene Discovery Through Functional Genomics. *Science*, 319(5863):620–624, Feb. 2008. ISSN 0036-8075, 1095-9203.
- [125] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 978-0-262-19475-4.
- [126] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.
- [127] E. Scornet. Random Forests and Kernel Methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, Mar. 2016. ISSN 0018-9448.
- [128] A. J. Sedgewick et al. Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM. *Bioinformatics*, 29(13):i62–i70, July 2013. ISSN 1367-4803.
- [129] J. A. Seoane, I. N. M. Day, T. R. Gaunt, and C. Campbell. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, 30(6):838–845, Mar. 2014. ISSN 1367-4803, 1460-2059.
- [130] D. D. Shao et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Research*, 23(4):665–678, Apr. 2013. ISSN 1088-9051.
- [131] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [132] R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, Nov. 2009. ISSN 1367-4803.
- [133] C. L. Shields et al. Prognosis of uveal melanoma in 500 cases using genetic testing of fine-needle aspiration biopsy specimens. *Ophthalmology*, 118(2):396–401, Feb. 2011. ISSN 1549-4713.
- [134] A. D. Singh, M. E. Turell, and A. K. Topham. Uveal melanoma: trends in incidence, treatment, and survival. *Ophthalmology*, 118(9):1881–1885, Sept. 2011. ISSN 1549-4713.
- [135] A. Sokolov et al. Pathway-Based Genomics Prediction using Generalized Elastic Net. *PLOS Computational Biology*, 12(3):e1004790, Mar. 2016. ISSN 1553-7358.

- [136] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, page 35, 2006.
- [137] G. Stolovitzky, D. Monroe, and A. Califano. Dialogue on Reverse-Engineering Assessment and Methods. *Annals of the New York Academy of Sciences*, 1115(1):1–22, 2007. ISSN 1749-6632.
- [138] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, Jan. 2007. ISSN 1471-2105.
- [139] A. Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, Oct. 2005.
- [140] P. H. Sudmant et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, Oct. 2015. ISSN 1476-4687.
- [141] R. J. Sullivan et al. First-in-Class ERK1/2 Inhibitor Ulixertinib (BVD-523) in Patients with MAPK Mutant Advanced Solid Tumors: Results of a Phase I Dose-Escalation and Expansion Study. *Cancer Discovery*, 8(2):184–195, 2018. ISSN 2159-8290.
- [142] Y. Sun et al. Androgen deprivation causes epithelial-mesenchymal transition in the prostate: implications for androgen-deprivation therapy. *Cancer Research*, 72(2):527–536, Jan. 2012. ISSN 1538-7445.
- [143] B. Sunkel et al. Integrative analysis identifies targetable CREB1/FoxA1 transcriptional co-regulation as a predictor of prostate cancer recurrence. *Nucleic Acids Research*, 44(9):4105–4122, 2016. ISSN 1362-4962.
- [144] T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, 41(3):1381–1405, June 2013. ISSN 0090-5364.
- [145] T. Suzuki and R. Tomioka. SpicyMKL: a fast algorithm for Multiple Kernel Learning with thousands of kernels. *Machine Learning*, 85(1-2):77–108, Oct. 2011. ISSN 0885-6125, 1573-0565.
- [146] X. TAN et al. Epidermal Growth Factor Receptor: A Novel Target of the Wnt/ β -Catenin Pathway in Liver. *Gastroenterology*, 129(1):285–302, July 2005. ISSN 0016-5085.

- [147] The HPN-DREAM Consortium et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318, Apr. 2016. ISSN 1548-7091, 1548-7105.
- [148] V. Thorsson et al. The Immune Landscape of Cancer. *Immunity*, 48(4): 812–830.e14, Apr. 2018. ISSN 1074-7613.
- [149] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
- [150] R. Tomioka and T. Suzuki. Sparsity-accuracy trade-off in MKL. *arXiv:1001.2615 [stat]*, Jan. 2010. arXiv: 1001.2615.
- [151] R. Tomioka, T. Suzuki, and M. Sugiyama. Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparsity Regularized Estimation. *Journal of Machine Learning Research*, 12:1537–1586, May 2011. ISSN 1533-7928.
- [152] F. Tschentscher et al. Tumor Classification Based on Gene Expression Profiling Shows That Uveal Melanomas with and without Monosomy 3 Represent Two Distinct Entities. *Cancer Research*, 63(10):2578–2584, May 2003. ISSN 0008-5472, 1538-7445.
- [153] Y.-Y. Tseng et al. PVT1 dependence in cancer with MYC copy-number increase. *Nature*, 512(7512):82–86, Aug. 2014. ISSN 0028-0836.
- [154] A. Tsherniak et al. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576.e16, July 2017. ISSN 0092-8674.
- [155] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, Apr. 2001. ISSN 0027-8424.
- [156] A. W. v. d. Vaart, S. Dudoit, and M. J. v. d. Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.
- [157] T. van den Bosch et al. Higher percentage of FISH-determined monosomy 3 and 8q amplification in uveal melanoma cells relate to poor patient prognosis. *Investigative Ophthalmology & Visual Science*, 53(6):2668–2674, May 2012. ISSN 1552-5783.

- [158] M. van der Laan and S. Dudoit. Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. Technical Report 130, University of California, Berkeley U.C. Berkeley Division of Biostatistics Working Paper Series, 2003.
- [159] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), Jan. 2007. ISSN 1544-6115, 2194-6302.
- [160] C. J. Vaske et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, June 2010. ISSN 1367-4803, 1460-2059.
- [161] A. Vellaichamy et al. Proteomic interrogation of androgen action in prostate cancer cells reveals roles of aminoacyl tRNA synthetases. *PloS One*, 4(9):e7075, Sept. 2009. ISSN 1932-6203.
- [162] M. Venkatesh et al. In vivo and in vitro characterization of a first-in-class novel azole analog that targets pregnane X receptor activation. *Molecular Pharmacology*, 80(1):124–135, July 2011. ISSN 1521-0111.
- [163] G. Virgili et al. Incidence of uveal melanoma in Europe. *Ophthalmology*, 114(12):2309–2315, Dec. 2007. ISSN 1549-4713.
- [164] Q. Wan and R. Pal. An Ensemble Based Top Performing Approach for NCI-DREAM Drug Sensitivity Prediction Challenge. *PLoS ONE*, 9(6), June 2014. ISSN 1932-6203.
- [165] H. Wang et al. Activated Pregnenolone X-Receptor Is a Target for Ketoconazole and Its Analogs. *Clinical Cancer Research*, 13(8):2488–2495, Apr. 2007. ISSN 1078-0432, 1557-3265.
- [166] Q. Wang et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell*, 138(2):245–256, July 2009. ISSN 1097-4172.
- [167] E. Weis et al. The association between host susceptibility factors and uveal melanoma: a meta-analysis. *Archives of Ophthalmology (Chicago, Ill.: 1960)*, 124(1):54–60, Jan. 2006. ISSN 0003-9950.
- [168] D. A. Wheeler and L. R. Roberts. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, 169(7):1327–1341.e23, June 2017. ISSN 0092-8674.

- [169] D. H. Wolpert. Stacked Generalization. *Neural Networks*, 5:241–259, 1992.
- [170] S. E. Woodman. Metastatic uveal melanoma: biology and emerging treatments. *Cancer Journal (Sudbury, Mass.)*, 18(2):148–152, Apr. 2012. ISSN 1540-336X.
- [171] M. N. Wright and A. Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, Mar. 2017. ISSN 1548-7660.
- [172] F. Xie et al. Identification of a Potent Inhibitor of CREB-Mediated Gene Transcription with Efficacious in Vivo Anticancer Activity. *Journal of Medicinal Chemistry*, 58(12):5075–5087, June 2015. ISSN 0022-2623.
- [173] S. Yavuzigitoglu et al. Uveal Melanomas with SF3b1 Mutations: A Distinct Subclass Associated with Late-Onset Metastases. *Ophthalmology*, 123(5):1118–1128, 2016. ISSN 1549-4713.
- [174] H. Yu et al. Tumor suppressor and deubiquitinase BAP1 promotes DNA double-strand break repair. *Proceedings of the National Academy of Sciences of the United States of America*, 111(1):285–290, Jan. 2014. ISSN 1091-6490.
- [175] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, Feb. 2006. ISSN 1369-7412, 1467-9868.
- [176] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, Apr. 2005. ISSN 1467-9868.
- [177] K. Zuberi et al. GeneMANIA Prediction Server 2013 Update. *Nucleic Acids Research*, 41(W1):W115–W122, July 2013. ISSN 0305-1048.